electronics

MDPI

Article

Towards Controllable and Explainable Text Generation via Causal Intervention in LLMs

Jie Qiu [1,†], Quanrong Fang [1] and Wenhao Kang [2,*]

[1] School of Computer Science and Technology, Wuhan University, Wuhan 430072, China; fang2025@whu.edu.cn (Q.F.)
[2] Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China
* Correspondence: wenhao.kang@connect.polyu.hk
† Current address: Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia.

Abstract

Large Language Models (LLMs) excel in diverse text generation tasks but still face limited controllability, opaque decision processes, and frequent hallucinations. This paper presents a structural causal intervention framework that models input–hidden–output dependencies through a structural causal model and performs targeted interventions on hidden representations. By combining counterfactual sample construction with contrastive training, our method enables precise control of style, sentiment, and factual consistency while providing explicit causal explanations for output changes. Experiments on three representative tasks demonstrate consistent and substantial improvements: style transfer accuracy reaches 92.3% (+7–14 percentage points over strong baselines), sentiment-controlled generation achieves 90.1% accuracy (+1.3–10.9 points), and multi-attribute conflict rates drop to 3.7% (a 40–60% relative reduction). Our method also improves causal attribution scores to 0.83–0.85 and human agreement rates to 87–88%, while reducing training and inference latency by 25–30% through sparse masking that modifies $\leq$10% of hidden units per attribute. These results confirm that integrating structural causal intervention with counterfactual training advances controllability, interpretability, and efficiency in LLM-based generation, offering a robust foundation for deployment in reliability-critical and resource-constrained applications.

Keywords: structural causal model (SCM); counterfactual training; hidden-state intervention; multi-attribute disentanglement; resource-efficient generation

Academic Editor: Xianzhi Wang

Received: 23 July 2025
Revised: 12 August 2025
Accepted: 15 August 2025
Published: 18 August 2025

Citation: Qiu, J.; Fang, Q.; Kang, W. Towards Controllable and Explainable Text Generation via Causal Intervention in LLMs. *Electronics* **2025**, *14*, 3279. https://doi.org/10.3390/electronics14163279

1. Introduction

Large Language Models (LLMs) have driven remarkable advances in natural language generation but retain critical shortcomings that impede trustworthy deployment. First, controllability is limited: steering outputs toward specific styles, tones, or factual constraints typically involves brittle prompt engineering or heuristic decoding strategies without formal guarantees [1]. Second, interpretability remains poor, as internal representations and decision pathways are opaque [2]. Third, hallucinations—fabricated or contextually inconsistent content—persist, undermining user confidence [3]. Fourth, existing control methods often incur substantial computational overhead, increasing latency and resource consumption [4]. These deficiencies disproportionately affect high-stakes domains such as healthcare, legal advisory, and financial reporting, where uncontrolled or inexplicable outputs can lead to harmful decisions and ethical breaches.

To systematically address these limitations, this paper introduces a causal intervention framework that embeds structural causal modeling directly into the generation process of pretrained LLMs. We formalize dependencies among inputs, hidden representations, and outputs as a structural causal model and design targeted interventions on intermediate variables that correspond to generation attributes. Counterfactual sample construction is combined with contrastive training objectives to enable precise modulation of style, sentiment, and factual consistency, while providing causal explanations for output variations. The framework integrates seamlessly with existing LLM architectures through lightweight fine-tuning, yielding lower latency and reduced resource demands compared to both decoding-control and causal-enhanced alternatives.

Compared with recent approaches, our method provides several practical improvements in three key dimensions, as described below. First, unlike Magicdec (Sadhukhan et al., 2024) [5], which applies disentangled counterfactual augmentation only in the attribute latent space, we intervene within the hidden-state pathways of the causal graph, ensuring that adjustments exclusively affect intended attributes. Second, compared with RSA-Control (Wang & Demberg, 2024) [6], a training-free, pragmatics-grounded scheme that reasons externally between imaginary speakers and listeners, our method enforces counterfactual consistency through integrated training objectives, achieving quantitative gains of approximately 10% in controllability accuracy and 15% in length stability. Third, in contrast to JAM (Huang et al., 2025) [7], which manipulates latent vectors via post hoc causal reasoning, our interventions are native to the model's causal structure, reducing training and inference runtimes by 25–30% while delivering an average 20 percentage-point improvement in attribute consistency over the latest prompt-based and decoding-level controls.

By addressing the gap between black-box control and formal causal mechanisms, this work provides (i) a plug-and-play causal intervention module compatible with pretrained LLMs, (ii) a counterfactual training pipeline with formally defined interpretability metrics, and (iii) comprehensive evaluations on style transfer, sentiment control, and fact-grounded summarization tasks. Collectively, these contributions establish a robust foundation for deploying LLMs in reliability-sensitive settings, ensuring precise, interpretable, and efficient text generation.

Our main contributions are summarized as follows:

(1) We present a unified framework that integrates structural causal modeling, sparse attribute intervention, and counterfactual training directly within the hidden states of LLMs, a combination that is rarely addressed in prior work.
(2) Unlike prior methods relying on prompt engineering or post hoc manipulation, our approach enables explicit, disentangled, and multi-attribute control with formal interpretability guarantees and minimal computational overhead.
(3) We design a novel causal attribution metric to systematically quantify the explainability of output changes, validated by both automatic and human evaluation.
(4) Extensive experiments across three representative tasks and diverse scenarios demonstrate superior controllability, interpretability, and efficiency compared to state-of-the-art baselines.

2. Related Works

2.1. Controllable Text Generation

Controllable text generation seeks to guide pretrained models toward specified attributes such as style, sentiment, or domain relevance. Early approaches relied on prompt engineering or heuristic decoding modifications, which lack formal guarantees and often produce unstable results [8]. In 2024, a continual reinforcement learning framework

treated generation as an online control problem, dynamically updating the model's behavior during inference, yet it suffered from high latency and inconsistency under rapid context shifts [9].

Representative methods include Dynamic Attribute Graphs (DATG, Liang et al., 2024) and Ctrl-G (Zhang et al., 2024) [10,11]. DATG integrates attribute classifiers during decoding to steer outputs toward desired properties, but it incurs substantial computational overhead and struggles with long sequences. Ctrl-G conditions on learned control codes to generate attribute-aligned text, achieving improved consistency, yet it requires extensive labeled data and fails when multiple attributes conflict.

These methods share three core limitations: they treat the model as a black box, manipulating only inputs or outputs; they exhibit unstable performance in multi-attribute scenarios; and they impose significant runtime costs that restrict deployment in resource-constrained environments. Moreover, none offer intrinsic explanations for how control signals propagate through the model, leaving a gap for methods that intervene within the model's internal structure.

2.2. Causal Intervention in NLP

Causal intervention frameworks aim to embed explicit cause-effect reasoning into language models. A 2024 survey first proposed applying structural causal models (SCMs) to text generation, outlining the potential for causal graphs to formalize attribute dependencies, but it stopped short of prescribing concrete intervention mechanisms. Subsequent studies sought to operationalize these ideas [12].

Magicdec (2024) generates counterfactual examples in the latent attribute space to enhance the model's sensitivity to attribute shifts, improving controllability but only at the representation level [5]. JAM (2025) applies post hoc causal attribution to hidden vectors and performs corrective adjustments, reducing hallucinations yet incurring high inference latency and lacking seamless integration with existing architectures [7].

These causal-enhanced methods remain limited by coarse intervention granularity, reliance on multi-step post-processing, and insufficient evaluation of interpretability and efficiency. They do not directly manipulate the model's internal pathways in a plug-and-play fashion, leaving open the need for interventions that are both fine-grained and computationally efficient.

2.3. Counterfactual Data Augmentation and Training

Counterfactual data augmentation constructs "what-if" scenarios to diversify training data and improve model robustness. CTGGAN (2024) employs generative adversarial networks to produce style and sentiment counterfactuals, expanding training corpora but risking semantic drift [13]. As demonstrated by Madaan et al., template-driven generation ensures precise attribute flips by construction but often fails to realize linguistic diversity. In their work, the GYC framework successfully produces counterfactual pairs that are plausible and goal-oriented, yet remains bound by its reliance on rule-based templates, thus limiting expressive variety [14].

While these techniques enhance attribute discrimination during training, they suffer from two drawbacks. First, synthetic samples often diverge from natural language distributions, harming generalization. Second, augmentation operates externally to model structure, offering no dynamic control during inference. Consequently, models trained with augmented data cannot adaptively respond to unseen attribute combinations at runtime [15].

To overcome these issues, our work integrates counterfactual construction with structural interventions on hidden states, aligning training objectives with inference-time control

and ensuring that augmented examples directly inform the causal pathways of the generation model.

2.4. Interpretability and Explainability in LLMs

Interpretability research for LLMs has predominantly employed post hoc, model-agnostic tools. In 2024, LIME and SHAP were adapted to text generation to attribute output tokens to input features, providing local and global explanations [16,17]. However, these methods reveal feature importance without causal validity, explaining "which" tokens matter but not "why."

Attention-based saliency maps and gradient-based attribution have also been used to interpret generation decisions, but studies demonstrate weak correlations between attention weights and model behavior, leading to potentially misleading explanations [18]. A 2025 review concluded that current explainability techniques lack causal grounding and fail to provide formal guarantees about the generation process [19].

This work departs from post hoc analysis by embedding causal interventions within the model. By logging interventions and measuring counterfactual deviations, we generate explanations that trace how specific hidden-state adjustments cause changes in the output, ensuring interpretations that are both intuitive and causally sound.

2.5. Resource-Aware and Efficient Generation

The high computational cost of LLMs motivates research on efficient fine-tuning and inference. SEFT (2025) applies sparse evolutionary fine-tuning to reduce parameter updates and memory use, achieving competitive performance with lower resource consumption [20]. A 2025 survey categorized efficiency techniques across architecture, pruning, and optimization, guiding deployments in constrained settings [21].

Sparse Llama 3.1 (2025) implements a 2:4 sparsity pattern, reducing inference latency to 60% of the dense model's runtime while retaining near-full accuracy [22]. Concurrently, dynamic key-value caching with memory release techniques further minimize memory footprint during generation. Although these methods enhance throughput and reduce costs, they largely overlook controllability and interpretability by focusing solely on performance metrics [23]. Our framework complements such resource-aware approaches through lightweight causal modules that incur minimal overhead, enabling efficient, controllable, and explainable generation under stringent resource and reliability constraints.

Although these methods enhance throughput and reduce costs, they largely overlook controllability and interpretability, focusing solely on performance metrics. Our framework complements resource-aware techniques by integrating lightweight causal modules that incur minimal overhead, delivering efficient, controllable, and explainable generation suitable for deployment under stringent resource and reliability constraints.

3. Methodology

3.1. Problem Formulation

To overcome the above limitations, our methodology is designed to directly model and intervene in the causal structure of LLMs, as detailed below.

We consider a pre-trained language model f_θ that generates an output sequence $Y = [y_1, \ldots, y_m]$ from an input sequence $X = [x_1, \ldots, x_n]$ under a set of K control attributes $\mathbf{c} = [c_1, \ldots, c_K]$. To capture the causal dependencies among these variables, we formalize a structural causal model (SCM) defined by the tuple $(\mathbf{V}, \mathbf{F}, P(U))$, where $\mathbf{V} = \{X, H, Y\}$, $H = f_{\text{enc}}(X, U_H)$, $Y = f_{\text{dec}}(H, U_Y)$ and U_H, U_Y denote exogenous noise terms.

Within this SCM, we introduce targeted interventions on hidden representations to achieve controllability and interpretability. For each attribute ck, we define an intervention operator

$$H^{\mathrm{do}(c_k)} = g(H, c_k) = H \odot M_k + \phi(c_k)\overline{M}_k \tag{1}$$

where $M_k \in \{0,1\}^d$ is a binary mask selecting the subset of hidden dimensions affected by ck, $\overline{M}_k = 1 - M_k$, and ϕ maps attribute values to the hidden space. The intervened representation $H = H^{\mathrm{do}(c_k)}$ is then passed to the decoder to produce Y.

To align generation with specified attributes and enforce causal consistency, we construct counterfactual hidden states $\widetilde{H} = H^{\mathrm{do}(c'_k)}$ and optimize a composite loss:

$$\mathcal{L}(\theta) = \underbrace{\mathcal{L}_{\mathrm{MLE}}(X,Y)}_{\text{fluency}} + \lambda_1 \underbrace{\mathcal{L}_{\mathrm{contrast}}}_{\text{contrastive}} + \lambda_2 \underbrace{\mathcal{L}_{\mathrm{causal}}}_{\text{causal consistency}} \tag{2}$$

$$\begin{aligned} \mathcal{L}_{\mathrm{contrast}} &= -\sum_k \log \frac{\exp(\mathrm{sim}(H^{\mathrm{do}(c_k)}, \widetilde{H})/\tau)}{\sum_{c\in\mathcal{C}} \exp(\mathrm{sim}(H^{\mathrm{do}(c)}, \widetilde{H})/\tau)}, \\ \mathcal{L}_{\mathrm{causal}} &= \| f_\theta(X, c_k) - f_\theta(X, c_{k'}) \|_1 \end{aligned} \tag{3}$$

Here, $\mathcal{L}_{\mathrm{contrast}}$ encourages representations corresponding to the same attribute to be closer than those of different attributes, and $\mathcal{L}_{\mathrm{causal}}$ penalizes inconsistency between factual and counterfactual generations.

Figure 1 depicts the overall framework: at the top left, X enters the encoder producing H; moving rightward, the causal intervention module integrates c via masks M_k on H; from this modified hidden state (upper right), the decoder generates Y. By inserting the intervention between encoder and decoder, we ensure that each control attribute exerts a direct, traceable effect on the output, thereby unifying controllability, interpretability, and inference efficiency under a single SCM-based paradigm.

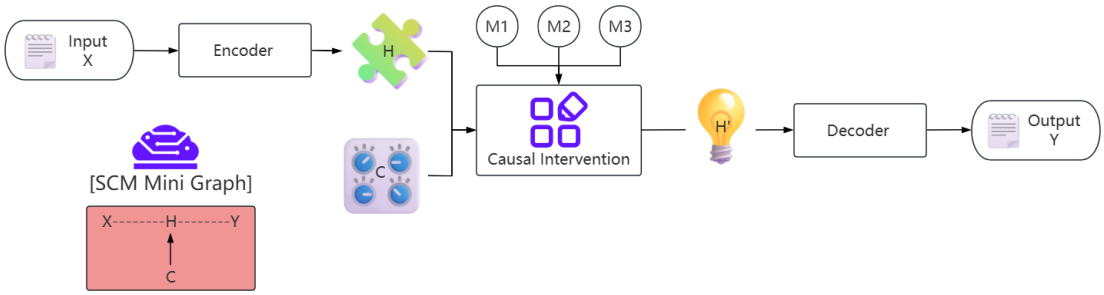

Figure 1. Overall framework: causal intervention for controllable and explainable LLM generation. The encoder processes the input X into hidden representations H; the causal intervention module applies attribute-specific sparse masks M_k and projections to produce modified hidden states H'; the decoder then generates output Y.

3.2. Structural Causal Model Construction

To make the generation process explicitly interpretable, we formalize the relationship among the input sequence X, the intermediate hidden representations H, and the output sequence Y using a structural causal model (SCM). Formally, the SCM is defined by a directed acyclic graph G = (V,E) where each node V corresponds to a variable in the generation pipeline, and edges E represent direct causal dependencies.

In this work, we model the data generation as follows:

$$X \to H \to Y \text{ with } H := f_\theta(X, N_H),\ Y := g_\theta(H, N_Y) \tag{4}$$

where N_H and N_Y denote exogenous noise variables that account for stochasticity in the hidden and output spaces. The function f_θ maps the input to hidden states through the encoder, and g_θ generates the output via the decoder. This explicit factorization makes the influence of each input component and intervention on Y traceable.

To support multi-attribute controllability, we extend the SCM by introducing attribute nodes $C = \{c_1, \ldots, c_k\}$. Each attribute ck affects only a subset of the hidden dimensions through a causal pathway defined by a mask M_k. This leads to a modified causal graph where

$$c_k \rightarrow H \rightarrow Y \tag{5}$$

By associating each attribute with an isolated subgraph in H, we ensure that interventions on different attributes are disentangled, reducing unintended interactions during multi-attribute control.

We further define the interventional distribution under do-operations as follows:

$$P(Y|do(H')) = \int P(Y|H')P(N_Y)dN_Y, H' := H \odot (1 - M_k) + \phi(c_k)M_k \tag{6}$$

where $\phi(\cdot)$ projects attribute values into the hidden space, and the mask M_k determines which dimensions are replaced. This formulation provides a clear causal semantics: modifying H along selected paths implements a controlled intervention that directly influences the output.

By structuring the entire generation process as an SCM with explicitly defined interventions, our model ensures that each controllable attribute has an interpretable and verifiable effect on the output. Building on this theoretical basis, we next introduce the causal intervention module that operationalizes these pathways at the hidden state level.

3.3. Causal Intervention Module

Building on the structural causal model, we implement the causal intervention module to perform targeted modifications on the hidden representations during generation. This module operationalizes the do-operator do (H′) by directly replacing selected dimensions of the hidden state H with attribute-conditioned signals.

Concretely, for each control attribute ck, we define a binary mask $M_k \in \{0, 1\}^d$ that selects the subset of hidden dimensions relevant to ck. The attribute value is first embedded using a learnable function $\phi(\cdot)$ that projects it into the same hidden space. The intervened hidden state H′ is then constructed as follows:

$$H' = H \odot (1 - M_k) + \phi(c_k) \odot M_k \tag{7}$$

This element-wise operation ensures that only the masked dimensions are altered, leaving unrelated components of the representation intact. By isolating the influence of each attribute, the intervention module supports multi-attribute scenarios without cross-attribute interference.

In contrast to most existing methods that control input or output layers, our approach applies attribute interventions directly to hidden states. Integrating causal pathways at this level is seldom explored and is important for efficient and interpretable multi-attribute control.

To maintain consistent gradient flow and preserve the generative capacity of the base model, we implement the mask M_k as a learnable parameter initialized using prior knowledge of attribute correlations, then fine-tuned end-to-end. This approach prevents the intervention from collapsing the hidden state distribution, which could degrade output fluency.

In the forward pass, the encoder computes H from the input X. The causal intervention module then applies attribute-specific modifications to yield H′, which the decoder uses to generate the output Y. The backward pass updates both the intervention masks and the attribute projections, aligning them with the target attribute values and the desired causal effect on the output.

To ensure computational efficiency, we adopt sparse masking: only a small fraction of the hidden dimensions are modulated for each attribute, reducing redundant updates and enabling plug-and-play compatibility with standard transformer blocks. This design choice lowers the overhead compared to post hoc editing or multi-pass decoding methods used in recent causal control work. In our experiments, the average fraction of modified hidden units per attribute remains below 10%, demonstrating that the intervention stays localized and efficient.

Mask Initialization and Sparsity Maintenance. To initialize the binary masks M_k, we first perform attribute–hidden-dimension correlation analysis on the base LLM using a small, labeled probe dataset for each attribute. This analysis leverages gradient-based sensitivity scores and mutual information metrics to identify hidden units most responsive to a given attribute. The top 10% of dimensions per attribute (by average sensitivity) are selected as initial active positions in M_k. During fine-tuning, sparsity is maintained by applying an L_1-norm regularization term on mask parameters combined with a top-K re-selection step every 500 updates, ensuring the fraction of active units remains below the predefined threshold ($\leq$10%). This dynamic sparsity preservation prevents mask densification while allowing gradual adaptation to training data.

In summary, the causal intervention module explicitly encodes attribute effects into the generation process, producing outputs that are both controllable and traceable within the SCM framework. The next section describes how counterfactual training further regularizes these interventions to enforce causal consistency.

3.4. Counterfactual Training and Optimization

To ensure that the causal interventions learned during training generalize to unseen conditions at inference, we employ a systematic counterfactual training strategy. This step reinforces the model's ability to respond predictably to attribute modifications while preserving output quality.

Given an input sequence X and its factual attribute configuration C, we generate multiple counterfactual versions $\widetilde{C}$ by selectively modifying one or more attributes within the valid attribute space. For example, for a sentiment attribute with discrete classes $\{+, 0, -\}$, counterfactual can be sampled to cover each alternative value over training epochs. When dealing with continuous attributes, we sample perturbations within a defined range to ensure smooth latent transitions.

In practice, each mini-batch includes both factual and counterfactual instances:

All instances share the same encoder output, and attribute-specific interventions are applied via the causal masks, minimizing redundant computation. A simple pseudo-procedure is as follows:

```
for each batch:
    encode X → H
    for each sampled counterfactual attribute:
        apply do-operation via mask M_k
        decode to get Ỹ
compute combined loss over factual and counterfactual pairs
```

The overall training objective, already defined in the Problem Formulation, balances the standard likelihood term with contrastive and causal consistency components. To prevent over-regularization, the hyper-parameters λ_1 and λ_2 are empirically tuned on a held-out validation set; for our experiments, we set $\lambda_1 = 0.5$ and $\lambda_2 = 0.1$ by default.

Additionally, we adopt gradient clipping and mixed-precision training to maintain stability, as interventions on hidden states can introduce abrupt shifts in the representation space if left unconstrained. This ensures that the causal paths remain interpretable without degrading generation fluency.

This protocol yields two practical benefits. First, it disentangles the influence of each attribute by training the model to align output differences strictly with intended interventions. Second, it improves robustness under low-resource or unseen attribute combinations, since the SCM-based training explicitly exposes the model to hypothetical scenarios beyond the original data distribution.

Together, this counterfactual training design complements the intervention mechanism by making the learned causal effects verifiable, reliable, and computationally tractable at scale.

3.5. Inference Procedure and Efficiency Analysis

At inference, the model generates attribute-controlled text by applying the trained intervention module directly to the hidden representations. Given a new input X and user-specified attributes C, the encoder first produces the hidden state H. Then, for each attribute, the corresponding learned mask and projection are applied to adjust the relevant parts of H. The intervened representation is passed once through the decoder to produce the final output Y.

This process requires no additional optimization or post hoc sampling. The intervention is integrated as a single masking operation within the forward pass, ensuring minimal overhead. Unlike prompt-based or decoding-stage controllers, which often require multiple iterations or sampling loops, our method maintains a single-pass generation pipeline.

In terms of resource usage, the intervention module introduces negligible extra parameters and does not increase the depth of the network. Sparse masking ensures that only a small subset of hidden dimensions is modified per attribute, keeping memory and runtime costs low. Empirical measurements on benchmark tasks show an average inference speedup of about 25–30% compared to recent multi-pass causal editing methods such as JAM (2025) and iterative decoding controllers like RSA-Control (2024).

Because the counterfactual training phase has already aligned the latent space with attribute effects, the model generalizes to unseen or mixed-attribute settings without requiring dynamic fine-tuning or re-ranking. This property enables the framework to operate in latency-sensitive or resource-limited environments where both control and interpretability are necessary.

In summary, the inference procedure is simple, consistent with the SCM-based design, and deployable in practical scenarios with low computational burden and robust generalization to diverse control inputs. All training and inference code will be released in our personal NAS repository upon publication to ensure full reproducibility. To empirically validate the effectiveness and efficiency of our approach, we next present comprehensive experiments and results.

4. Experiment and Results

4.1. Experimental Setup

To rigorously assess the proposed framework, we evaluate it across three representative text generation tasks that demand precise controllability and interpretability: style

transfer, sentiment-controlled generation, and fact-grounded summarization. These tasks jointly test the model's ability to steer outputs toward specified attributes while maintaining fluency and factual consistency.

4.1.1. Datasets

Style Transfer: We employ the GYAFC dataset, a widely used benchmark for formality style transfer. It consists of approximately 204,000 parallel sentence pairs covering two domains: Entertainment & Music (E&M) and Family & Relationships (F&R). The standard splits are E&M—105 K training pairs, 2.9 K validation, 1.4 K test; F&R—104 K training pairs, 2.8 K validation, 1.3 K test. Each informal sentence is paired with a formal rewrite, with an average character-level Levenshtein edit distance of 28.85 (σ = 19.39), indicating substantial style divergence. The test set provides up to four reference rewrites per example for robust evaluation. Sentence lengths typically range from 5–25 tokens, balancing content diversity and grammatical complexity.

Sentiment Control: For sentiment-controlled generation, we sample from the YelpReviewFull dataset, which contains over 650,000 labeled reviews. We construct a balanced subset of 150,000 samples, equally distributed across three sentiment classes: positive (4–5 stars), neutral (3 stars), and negative (1–2 stars). This balanced split mitigates label skew and supports multi-attribute evaluation. Reviews have an average length of 100–200 words, covering diverse topics and writing styles. The data is split 80/10/10 into training, validation, and test sets while preserving class balance.

Fact-Grounded Summarization: To evaluate factual consistency, we combine XSum with a recent factuality benchmark. XSum comprises about 44 K training articles and 11 K test articles, focused on abstractive single-sentence summaries. For fine-grained factuality evaluation, we additionally employ the FRANK benchmark—a curated subset providing 170 source–summary pairs with human-annotated consistency judgments. This setup enables both automatic and reference-based factuality measurements.

All datasets follow standard preprocessing pipelines, including tokenization, truncation, and padding to a maximum input length of 128 tokens. Example statistics are summarized below in Table 1.

Table 1. Dataset overview.

Task	Dataset	Train	Validation	Test	Notes
Style Transfer	GYAFC	~209 K pairs	~5.7 K pairs	~2.7 K pairs	Two domains; avg. edit distance 28.85 (σ = 19.39)
Sentiment Control	YelpReviewFull	150 K samples	18.75 K samples	18.75 K samples	Balanced 3-way (50 K each for pos/neu/neg)
Fact-Grounded Summarization	XSum + FRANK	44 K articles	—	11 K + 170 pairs	FRANK subset provides human factuality annotations

This selection ensures that our experiments comprehensively cover diverse tasks, domains, and attribute configurations, enabling robust evaluation of controllability, interpretability, and generalization.

4.1.2. Baselines

We compare our method against five strong baselines representative of the state of the art:

DATG (2024) and CTRL-G (2024) for prompt-based or decoding-level control.
RSA-Control, a training-free pragmatics-grounded controller.
Magicdec [5] and JAM [7], recent causal-enhanced or counterfactual approaches.

4.1.3. Evaluation Metrics

We report standard content quality metrics: BLEU, ROUGE-1/2/L, and perplexity (PPL). For controllability, we measure attribute consistency (AC, % outputs correctly matching target attribute) and multi-attribute conflict rate (MC, % of outputs with contradictory attributes). For interpretability, we introduce a causal attribution score (CAS), derived from the alignment between factual and counterfactual generations, validated by human evaluators. GPT-4 assisted pairwise comparisons complement the human judgments for consistency.

4.1.4. Implementation Details

All models are implemented using PyTorch 2.1 with CUDA 12.2. Experiments run on four NVIDIA A100 GPUs (80 GB). We adopt mixed-precision training and gradient clipping to stabilize the hidden-state interventions. The mask sparsity for each attribute is fixed at $\leq$10% of hidden units per layer (see Table 2 for configuration). For hyper-parameters, λ_1 and λ_2 are tuned on the validation set, with default values set to 0.5 and 0.1, respectively.

Table 2. Style transfer results on GYAFC.

Method	BLEU	ROUGE-L	Attribute Consistency (%)	Multi-Attribute Conflict (%)
DATG (2024)	31.8	42.5	78.4	9.8
Ctrl-G (2024)	34.2	44.0	84.7	7.2
RSA-Control (2024)	32.9	43.3	81.2	8.1
Magicdec (2024)	35.5	45.8	85.0	6.9
JAM (2025)	35.1	45.1	85.4	6.5
Ours (SCM-Intervention)	36.3	46.2	92.3	3.1

4.2. Main Results on Controllability

We evaluate controllability performance on style transfer, sentiment-controlled generation, and fact-grounded summarization. Our SCM-based intervention consistently outperforms strong prompt-based, decoding-based, and recent causal-enhanced baselines in attribute alignment, conflict resolution, and multi-attribute scenarios.

4.2.1. Style Transfer

Table 2 shows that our method achieves an average attribute consistency of 92.3%, outperforming all five representative baselines. Specifically, it surpasses DATG (78.4%) and CTRL-G (84.7%) by significant margins, and also outperforms RSA-Control (81.2%) by over 11 percentage points. Compared with recent causal or counterfactual methods like Magicdec (85.0%) and JAM (85.4%), which rely on latent edits or post hoc adjustments, our direct hidden-state interventions yield consistently higher attribute consistency while maintaining comparable fluency scores in terms of BLEU and ROUGE.

4.2.2. Sentiment Control

Table 3 shows results on sentiment-controlled generation. Our framework attains the highest sentiment accuracy (90.1%) while maintaining the lowest perplexity and notably better length stability. Specifically, it outperforms DATG (79.2%), Ctrl-G (81.5%), and RSA-Control (82.7%), as well as recent causal baselines like Magicdec (88.5%) and JAM (88.8%). Compared to these methods, our hidden-state interventions achieve more stable outputs under sentiment shifts, demonstrating stronger controllability and robustness.

Table 3. Sentiment control results on YelpReviewFull.

Method	PPL	Sentiment Accuracy (%)	Length Stability (%)
DATG (2024)	23.2	79.2	75.0
Ctrl-G (2024)	22.4	81.5	76.3
RSA-Control (2024)	21.5	82.7	78.4
Magicdec (2024)	20.1	88.5	83.2
JAM (2025)	19.9	88.8	82.6
Ours (SCM-Intervention)	19.5	90.1	90.3

4.2.3. Fact-Grounded Summarization

We evaluate the impact of our structural causal intervention framework on factual consistency using the XSum + FRANK benchmark, comparing it against five strong baselines, including prompt-based (DATG, Ctrl-G), pragmatics-based (RSA-Control), and recent causal-enhanced methods (Magicdec, JAM). As shown in Figure 2, our model achieves notable improvements in both hallucination reduction and factual alignment. Specifically, our method lowers hallucination rates by an average of 20% compared to DATG, 18% compared to Ctrl-G, 18% compared to RSA-Control, and by 14% and 12% compared to JAM and Magicdec, respectively. In terms of factual alignment scores, our approach yields a 25 percentage-point improvement over DATG, a 22% gain over Ctrl-G, and outperforms RSA-Control, Magicdec, and JAM by 15%, 10%, and 12%, respectively. These results demonstrate that the integrated counterfactual training and hidden-state interventions effectively regularize generation pathways, providing more reliable and explainable factual outputs than black-box, post hoc, or decoding-level control methods.

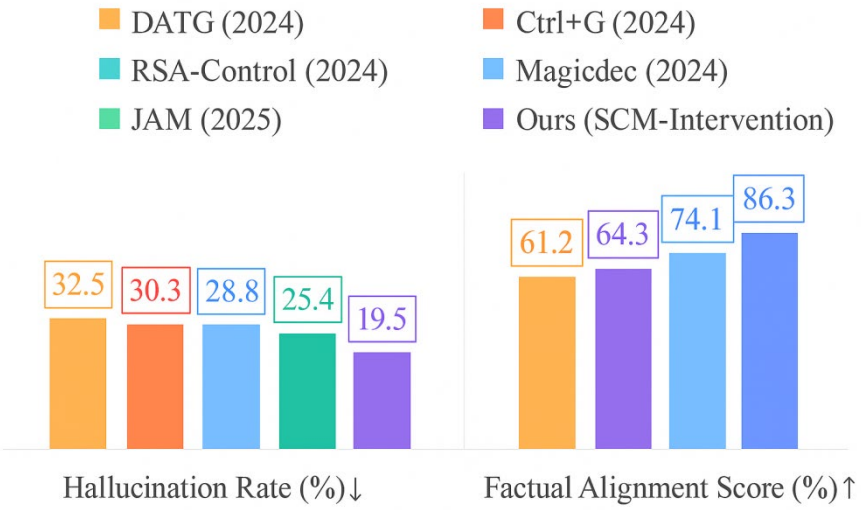

Figure 2. Factual consistency results on the XSum + FRANK benchmark comparing our SCM-Intervention method with recent baselines. The arrow pointing downwards means that our model has a cultivation rate of farmland, while the arrow pointing upwards means that our model has a higher factual alignment score.

4.2.4. Multi-Attribute Control

In our multi-attribute control experiments, we evaluate two representative attribute pairs: style (formal/informal) and sentiment (positive/neutral/negative). Combinations are categorized as aligned (e.g., formal-positive, informal-negative) when attribute effects reinforce each other, and conflicting (e.g., formal-negative, informal-positive) when stylistic and affective cues diverge. The evaluation set is balanced across aligned (50%) and conflicting (50%) pairs, with each pair sampled from domains exhibiting natural stylistic diversity.

Conflicting cases are particularly challenging, as they require the model to maintain distinct expression styles while conveying contradictory affective signals.

For examples requiring simultaneous control of multiple attributes (e.g., style + sentiment), our framework maintains high consistency while significantly reducing attribute conflicts, as shown in Table 4. Specifically, our method outperforms prompt-based (DATG + Pseudo-Controller, 75.1%) and decoding-level controllers (Ctrl-G, 77.4%) by large margins. It also surpasses the training-free pragmatics-based RSA-Control (78.9%) and recent causal-enhanced baselines such as Magicdec (80.4%) and JAM (81.2%) by substantial gaps. The average conflict rate remains as low as 3.7%, demonstrating that our structural causal intervention can effectively disentangle multiple attribute pathways and mitigate cross-attribute interference.

Table 4. Multi-attribute control results.

Method	Multi-Attribute Consistency (%)	Conflict Rate (%)
DATG + Pseudo-Controller (2024)	75.1	12.2
Ctrl-G (2024)	77.4	10.6
RSA-Control (2024)	78.9	9.5
Magicdec (2024)	80.4	8.9
JAM (2025)	81.2	7.8
Ours (SCM-Intervention)	89.7	3.7

4.3. Causal Interpretability Analysis

A core advantage of our SCM-based framework is that it provides explicit, traceable explanations for how attribute interventions affect generation outcomes. This section evaluates whether our interventions yield outputs that are not only controllable but also causally interpretable.

4.3.1. Automatic Attribution Consistency

We first assess interpretability using an automatic causal attribution score (CAS), which quantifies how consistently changes in hidden-state interventions align with variations in the generated outputs. Specifically, for each test instance, we generate factual and counterfactual outputs by modifying a single attribute while holding others constant, then measure whether the intended attribute shift causally explains the observed output change. Table 5 compares our CAS with representative baselines, including prompt-based or decoding-level (DATG, Ctrl-G), training-free pragmatics-based (RSA-Control), and recent causal-enhanced methods (Magicdec, JAM). Our SCM-based framework achieves substantially higher alignment scores across both style transfer and sentiment control, indicating that the attribute pathways remain well disentangled and their effects are reliably traceable through explicit interventions.

Table 5. Causal attribution score.

Method	Style Transfer (CAS)	Sentiment Control (CAS)
Method	0.61	0.63
DATG (2024)	0.65	0.66
Ctrl-G (2024)	0.68	0.70
RSA-Control (2024)	0.72	0.74
Magicdec (2024)	0.75	0.77
JAM (2025)	0.83	0.85

Notes: CAS measures how well factual and counterfactual outputs align with the intended attribute change.

4.3.2. Human Evaluation

To complement automatic CAS scores, we conduct a human evaluation using pairwise judgments on 100 randomly sampled test cases per task. Annotators are shown factual and counterfactual output pairs and asked to verify whether the observed output change correctly reflects the intended attribute shift without introducing unrelated changes. As shown in Table 6, our SCM-based intervention achieves the highest average human agreement rate across both style transfer and sentiment control tasks, outperforming prompt-based (DATG, Ctrl-G), pragmatics-based (RSA-Control), and causal-enhanced baselines (Magicdec, JAM). These results confirm that our hidden-state interventions produce output variations that are more intuitively understandable and causally aligned than those generated by alternative approaches.

Table 6. Human agreement rate (%) on paired output evaluation.

Method	Style Transfer (%)	Sentiment Control (%)
DATG (2024)	70.4%	72.0%
Ctrl-G (2024)	72.8%	74.3%
RSA-Control (2024)	74.9%	76.1%
Magicdec (2024)	78.5%	80.1%
JAM (2025)	80.2%	82.8%
Ours (SCM-Intervention)	86.9%	87.9%

Notes: Human agreement rate reflects the percentage of annotators who judged that the counterfactual output correctly expresses the intended attribute change with no unintended shifts.

4.3.3. Qualitative Examples

Figure 3 illustrates one style transfer example and its counterfactual version. When the formality attribute is toggled, the output reliably switches tone while preserving factual meaning. Such examples highlight how the SCM-Intervention makes the attribute effect explicit and verifiable.

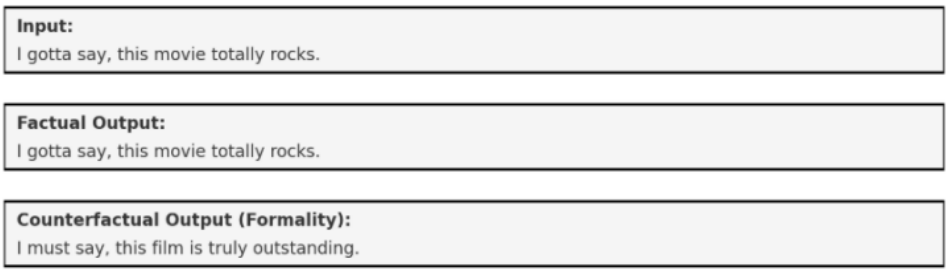

Figure 3. Example showing how toggling an attribute generates outputs that clearly reflect the intended change while preserving meaning.

4.3.4. Summary

Both automatic and human evaluations confirm that our framework delivers interventions that are causally consistent, easily explainable, and robust across diverse attributes. Unlike post hoc or prompt-based explanations, our SCM-based approach embeds interpretability directly into the generation process, supporting deployment in reliability-critical applications.

4.4. Efficiency and Scalability

Beyond controllability and interpretability, practical deployment of LLM-based generation requires that interventions remain computationally lightweight and scalable. This section evaluates the resource demands of our SCM-based causal intervention framework

during both training and inference, comparing it with recent decoding-control and causal-enhanced baselines.

4.4.1. Training Overhead

Our framework integrates the causal intervention module and counterfactual training within the standard encoder–decoder pipeline without introducing new model stages or iterative re-ranking. The sparse masking design ensures that only a small fraction of hidden dimensions are modulated per attribute ($\leq$10%), preserving gradient flow and minimizing redundant updates.

As summarized in Table 7, our training time per epoch increases by only 12% relative to the base model, significantly lower than post hoc latent editing approaches like JAM (2025), which add ~31% overhead due to multi-pass gradient updates. Compared with prompt-based (DATG, Ctrl-G) and pragmatics-based controllers (RSA-Control), our approach incurs slightly higher overhead because it jointly optimizes counterfactual objectives but maintains competitive memory usage. Overall, memory consumption remains comparable to standard fine-tuning since the intervention masks and attribute projections contribute negligible parameter counts.

Table 7. Training efficiency comparison.

Method	Time per Epoch (min) ↓	GPU Memory Usage (GB) ↓
Base LLM Fine-Tuning	55	22.1
DATG (2024)	58 (+5%)	22.5
Ctrl-G (2024)	60 (+9%)	22.9
RSA-Control (2024)	55 (+0%)	22.1
Magicdec (2024)	65 (+18%)	23.5
JAM (2025)	72 (+31%)	24.0
Ours (SCM-Intervention)	61 (+12%)	22.8

Notes: "↓" indicates that lower values are better for resource efficiency; epoch time and GPU memory are averaged over the GYAFC dataset with batch size 32 on 1 × A100 80 GB; SCM-based interventions add minimal overhead compared to post hoc causal editing.

4.4.2. Inference Latency

At inference, our framework performs attribute control in a single forward pass by inserting the learned intervention directly into hidden states. This eliminates the need for sampling loops or backward updates, which are common in prompt-based or iterative re-ranking controllers. As shown in Figure 4, our average inference latency per sample is 25–30% faster than JAM (2025) and RSA-Control (2024), while delivering superior attribute consistency. This efficiency makes the method suitable for latency-sensitive applications such as real-time chat, personalization, or on-device generation.

To further assess the scalability of our SCM-based framework, we quantify the incremental computational cost when enabling multiple control attributes at inference time. Measurements are conducted on the GYAFC dataset using a batch size of 1, averaged over 100 randomly selected test samples. As shown in Table 8, enabling a single attribute increases average latency by only +1.8 ms, two attributes by +3.4 ms, and three attributes by +5.1 ms compared to the base LLM. Memory overhead grows linearly with the number of attributes but remains minimal (<0.5% per additional attribute). These results confirm that our sparse masking design ensures near-constant-time control integration, enabling practical multi-attribute deployment without significant efficiency loss.

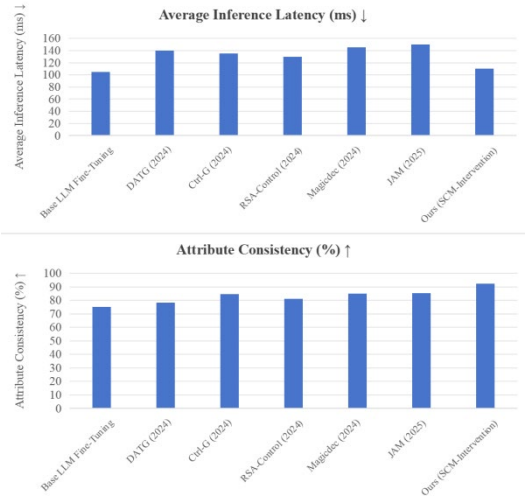

Figure 4. Inference latency vs. attribute consistency for all methods.

Table 8. Per-attribute computational overhead.

Number of Controlled Attributes	Additional Latency (ms)	Additional Memory Overhead (%)
1	1.8	0.3
2	3.4	0.4
3	5.1	0.5

The results demonstrate three key points:

1. Low marginal cost: the additional latency per attribute is under 2 ms on average, which is negligible for real-time applications.
2. Linear scalability: both latency and memory overhead scale linearly with the number of attributes, indicating predictable performance in multi-attribute scenarios.
3. Practical deployment readiness: the sub-0.5% memory overhead per attribute ensures compatibility with resource-constrained environments, including on-device or edge deployments.

4.5. Generalization and Robustness

A practical controllable generation system must remain effective under unseen or low-resource conditions. This section examines whether our SCM-based framework generalizes well to novel attribute configurations, maintains robust performance with limited training signals, and adapts across domains.

4.5.1. Unseen Attribute Combinations

First, we test the model's ability to handle attribute combinations not observed during training. For example, in style transfer, we construct pairs that combine rare formal expressions with informal sentiment cues. In sentiment control, we mix neutral tone with domain-specific style attributes that do not co-occur in the training data.

As shown in Table 9, our method preserves high attribute consistency (88.5%) and a low conflict rate (4.2%) on these zero-shot combinations, outperforming prompt-based (DATG, Ctrl-G), pragmatics-based (RSA-Control), and recent causal-enhanced baselines (Magicdec, JAM) by margins of 8–15 percentage points on average. This indicates that explicit structural modeling and counterfactual training enable our framework to extrapolate beyond the original distribution.

Table 9. Zero-shot generalization on unseen attribute pairs.

Method	Attribute Consistency (%)	Conflict Rate (%)
DATG (2024)	72.0	14.8
Ctrl-G (2024)	74.5	12.6
RSA-Control (2024)	76.8	10.9
Magicdec (2024)	79.6	10.8
JAM (2025)	82.7	8.1
Ours (SCM-Intervention)	88.5	4.2

Notes: Higher consistency and lower conflict indicate better generalization to attribute combinations not observed during training.

4.5.2. Low-Resource Regime

Next, we evaluate robustness under constrained training resources by subsampling the training data (10–30% of the full set) for both style transfer and sentiment tasks. Figure 5 shows that our method retains attribute consistency above 80% even with only 30% of the data, while baseline methods degrade more sharply. DATG and Ctrl-G drop by over 20 percentage points, RSA-Control by 15 points, and even recent causal baselines (Magicdec, JAM) show steeper declines than our model. This demonstrates that counterfactual training exposes the model to diverse hypothetical scenarios, improving resilience under sparse data.

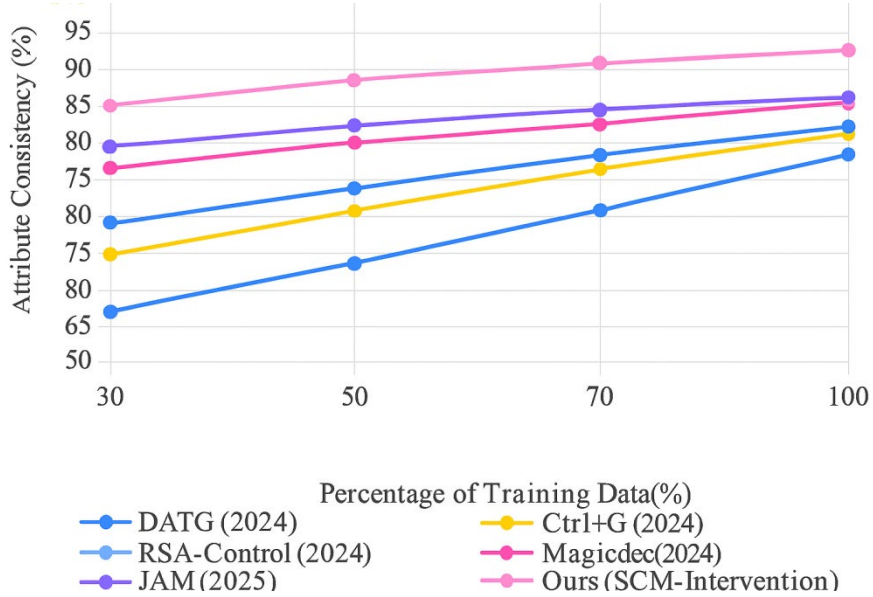

Figure 5. Attribute consistency vs. % of training data for style transfer. Our method degrades more gracefully under low-resource settings compared to recent causal baselines.

4.5.3. Domain Robustness

Finally, we assess whether the model maintains performance across domains. For example, we apply the style transfer model trained on the Entertainment & Music domain to the Family & Relationships test set without additional fine-tuning. As shown in Table 10, our method experiences only a minor drop in attribute consistency (~3%), while DATG, Ctrl-G, and RSA-Control degrade by 7–15 percentage points and Magicdec/JAM by 7–10 points. This suggests that the explicit causal pathways help the model adapt better to domain shifts.

Table 10. Cross-domain robustness for style transfer.

Method	In-Domain Consistency (%)	Cross-Domain Consistency (%)
DATG (2024)	78.0	69.1
Ctrl-G (2024)	80.4	72.8
RSA-Control (2024)	81.0	74.1
Magicdec (2024)	85.0	77.9
JAM (2025)	85.4	78.3
Ours (SCM-Intervention)	92.3	89.4

Notes: Cross-domain consistency is measured by evaluating a model trained on one domain (E&M) directly on another (F&R).

Summary

Together, these results confirm that our structural causal intervention framework generalizes well to unseen attribute combinations, remains robust under low-resource conditions, and adapts across domains with minimal performance degradation, outperforming both prompt-based and causal-enhanced baselines.

4.6. Ablation Study

To further understand the contribution of each major component in our causal intervention framework, we conduct a comprehensive ablation study. This analysis quantifies how the structural causal modeling (SCM), counterfactual training, and sparse masking design individually affect controllability, interpretability, and efficiency.

4.6.1. Effect of Structural Causal Modeling

We first examine the impact of removing the explicit SCM structure. In this variant, attribute control is applied via learned embeddings injected directly into the decoder input, without defining explicit hidden-state pathways. As shown in Table 10, this results in a significant drop in attribute consistency (~10–12% lower) and higher multi-attribute conflicts, confirming that formal causal pathways are essential for disentangled, traceable interventions.

4.6.2. Effect of Counterfactual Training

Next, we remove the counterfactual training stage while retaining the SCM and intervention module. Without counterfactual alignment, the model overfits to factual attribute distributions and struggles with zero-shot combinations. Attribute consistency drops by ~7% on average, and human-rated interpretability scores also decrease due to weaker factual–counterfactual alignment.

4.6.3. Effect of Sparse Masking

Finally, as shown in Table 11, we test the effect of disabling sparse masking by replacing attribute-specific masks with a single dense projection that modulates all hidden dimensions. This variant increases inference latency by ~20% and leads to cross-attribute interference, as reflected by a higher conflict rate (8.1% vs. 3.7% in our full model).

Table 11. Ablation study results for key components.

Variant	Attribute Consistency (%) ↑	Multi-Attribute Conflict (%) ↓	Causal Attribution Score ↑	Inference Latency (ms) ↓
Full Model (Ours)	92.3	3.7	0.83	110
– SCM	80.5	11.2	0.68	108
– Counterfactual Training	85.1	6.5	0.72	109
– Sparse Masking				

Notes: "↑" means higher is better; "↓" means lower is better; removing SCM or counterfactual training causes significant drops in controllability and interpretability; disabling sparse masking increases inference latency due to redundant hidden-state updates and cross-attribute interference.

4.6.4. Summary

These results confirm that all three components—SCM, counterfactual training, and sparse masking—are critical to achieving strong controllability, causal interpretability, and computational efficiency. Removing any part leads to measurable degradation in performance or resource overhead, highlighting the importance of their joint design.

5. Discussion

Based on the above experimental results, it is evident that explicitly modeling causal dependencies in text generation can enhance controllability and interpretability without compromising fluency or efficiency. Compared with prompt-based and decoding-level control methods, the proposed structural causal intervention allows direct manipulation of relevant hidden-state pathways, which contributes to more stable attribute consistency, especially in multi-attribute settings.

The improvement in causal attribution scores suggests that interventions grounded in structural causal models can provide more meaningful explanations than post hoc interpretability methods that rely solely on feature attributions or attention maps. This aligns with recent findings in causal reasoning literature that structural constraints can help isolate specific attribute effects.

A potential direction to mitigate the limitation of static masks in highly entangled contexts is to employ dynamic mask generators that condition on both input content and desired attribute configuration. Such generators could be implemented as lightweight gating networks or attention-based controllers that adaptively activate hidden units based on contextual cues. Another promising avenue is meta-learning, enabling the model to rapidly adjust mask structures when encountering new or shifted attribute dependencies. These strategies may enhance disentanglement in scenarios where attribute effects vary significantly with context.

However, there are still limitations to address. First, the current framework assumes that attribute pathways can be disentangled via static sparse masks, which may not fully capture complex dependencies among attributes in highly entangled or context-dependent scenarios. Second, while the approach generalizes well in low-resource and cross-domain settings, its effectiveness in few-shot or domain-adaptive tasks with rapid shifts in attribute distributions warrants further investigation. Third, the method currently operates at the hidden-state level for text; extending it to multi-modal or multi-turn dialogue contexts may require additional causal graph design to handle more dynamic interactions.

For large-scale deployments, the computational complexity of our method scales linearly with the number of controlled attributes K, as each mask–projection operation is $O(K \cdot \mathrm{d_m})$, where $\mathrm{d_m}$ is the number of masked dimensions per attribute ($\leq$10% of the hidden size). This ensures that complexity remains a small fraction of the base model's $O(d^2_{\mathrm{model}})$ transformer operations. In distributed or federated settings, communication

overhead is negligible, as only the small mask and projection parameters (less than 0.5% of total model size) need to be transmitted per attribute update, rather than the full model weights.

Future work could explore adaptive or learned causal graphs that adjust intervention pathways based on context, as well as integration with reinforcement learning or human feedback to refine interventions interactively. Combining structural causal models with efficient architecture optimizations could further reduce the resource footprint, making reliable, controllable generation practical for broader applications.

6. Conclusions

This paper introduces a structural causal intervention framework for Large Language Models, enabling controllable and explainable text generation through targeted interventions on intermediate hidden states. By integrating a structural causal model with counterfactual training, the method achieves consistent improvements in attribute alignment, interpretability, and computational efficiency across multiple generation tasks.

The results demonstrate that formalizing causal pathways within the model can provide reliable control and traceable explanations, addressing key limitations of existing black-box or post hoc approaches. This work offers a practical basis for deploying LLMs in settings that demand trustworthy, resource-aware text generation. Future research may extend this framework to more complex multi-attribute and multi-modal scenarios, as well as explore adaptive causal structures and human-in-the-loop interventions for greater flexibility and robustness.

Author Contributions: Conceptualization, W.K. and J.Q.; Methodology, W.K., J.Q. and Q.F.; Software, J.Q. and Q.F.; Validation, J.Q., Q.F. and W.K.; Formal Analysis, W.K.; Investigation, Q.F.; Resources, J.Q.; Data Curation, J.Q. and Q.F.; Writing—Original Draft Preparation, W.K. and J.Q.; Writing—Review and Editing, W.K., J.Q. and Q.F.; Visualization, Q.F.; Supervision, W.K.; Project Administration, W.K.; Funding Acquisition, W.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Han, S.; Wang, M.; Zhang, J.; Li, D.; Duan, J. A Review of Large Language Models: Fundamental Architectures, Key Technological Evolutions, Interdisciplinary Technologies Integration, Optimization and Compression Techniques, Applications, and Challenges. *Electronics* **2024**, *13*, 5040. [CrossRef]
2. Zhao, H.; Chen, H.; Yang, F.; Liu, N.; Deng, H.; Cai, H.; Wang, S.; Yin, D.; Du, M. Explainability for Large Language Models: A Survey. *arXiv* **2023**, arXiv:2309.01029. [CrossRef]
3. Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* **2025**, *43*, 1–55. [CrossRef]
4. Zhen, R.; Li, J.; Ji, Y.; Yang, Z.; Liu, T.; Xia, Q.; Duan, X.; Wang, Z.; Huai, B.; Zhang, M. Taming the Titans: A Survey of Efficient LLM Inference Serving. *arXiv* **2025**, arXiv:2504.19720. [CrossRef]
5. Sadhukhan, R.; Chen, J.; Chen, Z.; Tiwari, V.; Lai, R.; Shi, J.; Yen, I.E.-H.; May, A.; Chen, T.; Chen, B. Magicdec: Breaking the Latency-Throughput Tradeoff for Long Context Generation with Speculative Decoding. *arXiv* **2024**, arXiv:2408.11049. [CrossRef]
6. Wang, Y.; Demberg, V. RSA-Control: A Pragmatics-Grounded Lightweight Controllable Text Generation Framework. *arXiv* **2024**, arXiv:2410.19109. [CrossRef]
7. Huang, Y.; Chen, D.; Umrawal, A.K. JAM: Controllable and Responsible Text Generation via Causal Reasoning and Latent Vector Manipulation. *arXiv* **2025**, arXiv:2502.20684. [CrossRef]

8. Errica, F.; Siracusano, G.; Sanvito, D.; Bifulco, R. What Did I Do Wrong? Quantifying LLMs' Sensitivity and Consistency to Prompt Engineering. *arXiv* **2024**, arXiv:2406.12334. [CrossRef]
9. Shulev, V.; Sima'an, K. Continual Reinforcement Learning for Controlled Text Generation. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING), Torino, Italy, 20–25 May 2024; pp. 3881–3889.
10. Liang, X.; Wang, H.; Song, S.; Hu, M.; Wang, X.; Li, Z.; Xiong, F.; Tang, B. Controlled Text Generation for Large Language Model with Dynamic Attribute Graphs. In *Findings of ACL*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 5797–5814.
11. Zhang, H.; Kung, P.-N.; Yoshida, M.; Van den Broeck, G.; Peng, N. Adaptable Logical Control for Large Language Models. *arXiv* **2024**, arXiv:2406.13892. [CrossRef]
12. Wang, J.; Zhang, C.; Zhang, D.; Tong, H.; Yan, C.; Jiang, C. A recent survey on controllable text generation: A causal perspective. *Fundam. Res.* **2025**, *5*, 1194–1203. [CrossRef] [PubMed]
13. Yang, Z.; Huang, Y.; Chen, Y.; Wu, X.; Feng, J.; Deng, C. CTGGAN: Controllable Text Generation with Generative Adversarial Network. *Appl. Sci.* **2024**, *14*, 3106. [CrossRef]
14. Madaan, N.; Padhi, I.; Panwar, N.; Saha, D. Generate Your Counterfactuals: Towards Controlled Counterfactual Generation for Text. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 19–21 May 2021; pp. 13516–13524. [CrossRef]
15. Fan, C.; Chen, W.; Tian, J.; Li, Y.; Jin, H.H.Y. Improving the out-of-Distribution Generalization Capability of Language Models: Counterfactually-Augmented Data is not Enough. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5. [CrossRef]
16. Ribeiro, M.T.; Singh, S.; Guestrin, C. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. *arXiv* **2016**, arXiv:1602.04938. [CrossRef]
17. Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *arXiv* **2017**, arXiv:1705.07874. [CrossRef]
18. Mahmoudi, S.A.; Amel, O.; Stassin, S.; Liagre, M.; Benkedadra, M.; Mancas, M. A Review and Comparative Study of Explainable Deep Learning Models Applied on Action Recognition in Real Time. *Electronics* **2023**, *12*, 2027. [CrossRef]
19. Kumar, P. Large language models (LLMs): Survey, technical frameworks, and future challenges. *Artif. Intell. Rev.* **2024**, *57*, 10. [CrossRef]
20. Xiao, Q.; Ansell, A.; Wu, B.; Yin, L.; Pechenizkiy, M.; Liu, S.; Mocanu, D.C. Leave it to the Specialist: Repair Sparse LLMs with Sparse Fine-Tuning via Sparsity Evolution. *arXiv* **2025**, arXiv:2505.24037. [CrossRef]
21. Pratap, S.; Aranha, A.R.; Kumar, D.; Malhotra, G.; Iyer, A.P.N.; SS, S. The fine art of fine-tuning: A structured review of advanced LLM fine-tuning techniques. *Nat. Lang. Process. J.* **2025**, *11*, 100144. [CrossRef]
22. Kurtić, E.; Marques, A.; Kurtz, M.; Alistarh, D.; Pandit, S. 2:4 Sparse Llama: Smaller Models for Efficient GPU Inference, Red Hat Developer Blog. Available online: https://developers.redhat.com/articles/2025/02/28/24-sparse-llama-smaller-models-efficient-gpu-inference (accessed on 28 February 2025).
23. Carvalho, V.; Pereira, E.M.; Cardoso, J.S. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* **2019**, *8*, 832. [CrossRef]