CausalRoute: Causal Path Tracing and Hallucination Suppression in Multimodal Foundation Models

Brevor Qrtega
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
trevorortega97@buffalo.edu

Logan C. Graham
Department of Computer Science, University of North Texas, Denton, TX, USA.
contactlogan@unt.edu

Hugo Howard
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
hellohugo@oregonstate.edu

Abstract

Multimodal foundation models have demonstrated remarkable capabilities in integrating vision, language, and other sensory modalities, yet they remain susceptible to hallucinations—factually incorrect or contextually inconsistent outputs that undermine trust and reliability. Existing mitigation strategies, such as post-hoc verification and adversarial training, offer limited interpretability and fail to address the underlying causal mechanisms that generate erroneous outputs. This paper introduces CausalRoute, a framework for causal path tracing and hallucination suppression in multimodal foundation models. CausalRoute leverages structural causal models to trace the flow of information across modalities and identifies causal pathways that lead to hallucinated content. By performing targeted interventions along these pathways in the latent representation space, the method suppresses spurious correlations and enhances the factual grounding of generated outputs. We present a systematic analysis of the architectural trade-offs involved in integrating causal inference into large-scale multimodal systems, including computational overhead, scalability, and the interplay between causal interventions and model expressiveness. The framework is situated within broader considerations of infrastructure deployment, governance, and sustainability, emphasizing the need for interpretable and auditable AI systems. We discuss implications for fairness and robustness, particularly in high-stakes domains such as medical imaging and autonomous navigation, where hallucinations carry significant ethical and operational consequences. Policy perspectives are examined with respect to regulatory frameworks that demand transparency and accountability in AI-driven decision-making. CausalRoute represents a step toward ensuring that multimodal foundation models are not only powerful but also aligned with human expectations and safe for real-world deployment. The paper concludes with a forward-looking discussion on the integration of causal reasoning into the training and inference pipelines of next-generation multimodal systems.

Keywords

multimodal foundation models, causal inference, hallucination suppression, interpretability, system governance, robustness, fairness, policy.

1. Introduction

The rapid advancement of multimodal foundation models has reshaped the landscape of artificial intelligence, enabling systems to process and generate content across text, images, audio, and video with unprecedented fluency [1,2]. Models such as CLIP, DALL-E, and GPT-4V have demonstrated the ability to perform complex tasks ranging from visual question answering to image captioning and cross-modal retrieval [3,4]. However, a persistent challenge that undermines the practical deployment of these models is the phenomenon of hallucinations—outputs that are syntactically plausible but factually incorrect or inconsistent with the input modalities [5,6]. In visual question answering, for example, a model might describe a non-existent object in an image or incorrectly attribute a color to an object that is not present. Such errors erode user trust and pose significant risks in high-stakes applications such as medical diagnostics, autonomous driving, and legal document analysis.

Current approaches to hallucination suppression have largely focused on post-hoc correction mechanisms, such as retrieval-augmented generation, contrastive decoding, and self-consistency checks [7,8]. While these methods can reduce the frequency of errors, they often operate at the output level and do not address the root causal mechanisms that originate within the model's latent representations. Moreover, they provide limited insight into why a hallucination occurred, making it difficult to systematically improve model architecture or training procedures. There is a growing recognition that mitigating hallucinations requires a deeper understanding of how information flows through the multimodal processing pipeline and which causal pathways are responsible for generating erroneous content [9,10].

This paper introduces CausalRoute, a framework that combines causal path tracing with targeted intervention to suppress hallucinations in multimodal foundation models. The core idea is to construct a structural causal model of the multimodal inference process, identify the latent variables that mediate between input modalities and output tokens, and trace the causal paths along which spurious correlations propagate. By intervening on specific nodes along these paths—either by reweighting latent activations or by applying counterfactual perturbations—CausalRoute reduces the influence of confounding factors that lead to hallucinations. The approach is grounded in the principles of causal inference, which provide a rigorous mathematical foundation for distinguishing correlation from causation [11].

Beyond the technical details of the framework, this paper examines the broader system-level implications of integrating causal reasoning into large-scale multimodal systems. We discuss architectural trade-offs between the granularity of causal tracing and computational efficiency, the challenges of deploying such methods in resource-constrained environments, and the governance structures needed to ensure that causal interventions do not introduce new biases or reduce model expressiveness. We also consider the sustainability of causal inference methods in terms of energy consumption and hardware requirements. Throughout the paper, we emphasize the importance of aligning technical solutions with societal values, particularly fairness, accountability, and transparency. The discussion is framed within the context of emerging regulatory frameworks, such as the European Union's AI Act, that demand explainability and robustness in high-risk AI systems.

2. Background and Related Work

Multimodal foundation models represent a convergence of natural language processing and computer vision, enabled by large-scale pretraining on diverse datasets [1,2,3]. Architectures such as the transformer encoder-decoder, with cross-attention layers that fuse visual and textual features, form the backbone of these systems [4]. Despite their impressive performance, they are prone to hallucinations because the training objective—typically next-

token prediction or contrastive learning—does not explicitly penalize outputs that are inconsistent with the input modalities [5,6]. Hallucinations can arise from spurious correlations in the training data, from the model's reliance on language priors rather than visual evidence, or from insufficient representation of rare objects or interactions [7,8].

Existing methods for hallucination suppression fall into several categories. Post-hoc verification approaches use external knowledge bases or auxiliary models to check the factual consistency of generated outputs [9,10]. For instance, retrieval-augmented generation retrieves relevant documents or images and conditions the generation process on that information, thereby reducing the likelihood of generating unsupported claims [11]. Another class of methods modifies the decoding strategy, such as by penalizing the model when it assigns high probability to tokens that are inconsistent with the visual modality [12]. Contrastive decoding amplifies the difference between the model's own distribution and a weakened version to encourage factual predictions [13]. While these techniques have shown empirical success, they are fundamentally reactive and do not provide a causal understanding of why the model hallucinates.

Causal inference offers a principled alternative by shifting the focus from correlation to causation [14,15]. In the context of neural networks, causal tracing techniques have been developed to identify which layers or neurons are responsible for specific predictions [16]. These methods typically involve perturbing activations and observing changes in output probabilities, enabling the construction of causal graphs that map information flow. Work on causal abstraction in transformers has shown that intermediate representations encode semantically meaningful concepts and that interventions on those representations can alter model behavior in predictable ways [17]. However, most of this research has been confined to unimodal settings or small-scale models. Extending causal tracing to multimodal scenarios introduces new challenges because the causal pathways involve interactions between different modalities, each with its own representational structure and semantic granularity.

The concept of path-level intervention, as opposed to node-level intervention, has recently been proposed as a more robust way to modify model behavior [7]. Rather than intervening on a single activation, path-level intervention adjusts the entire causal pathway from input to output by reweighting the contributions of multiple intermediate representations. This approach is particularly suitable for multimodal models because hallucinations often involve a chain of errors that propagate through cross-modal attention layers. The framework presented in this paper builds on these ideas by designing a causal graph that explicitly models the relationships between visual features, textual embeddings, and the final generated tokens, and by implementing interventions that selectively suppress spurious pathways.

3. Causal Path Tracing: A Framework for Interpretability

Causal path tracing in multimodal foundation models begins with the construction of a directed acyclic graph that represents the inference process. In a typical transformer-based multimodal architecture, the input consists of a sequence of visual patches and textual tokens, each mapped to continuous embeddings. These embeddings pass through a series of transformer layers, where self-attention and cross-attention mechanisms compute interactions between modalities. The output of the final layer is decoded into a probability distribution over the next token. To trace causal paths, we define latent variables at each layer corresponding to the aggregated representations for visual and textual modalities. The causal graph includes edges that represent the flow of information from input to these latent variables and from the latent variables to the output.

The process of causal path tracing involves two main steps. First, we intervene on the input by replacing or corrupting a portion of the visual or textual cues, and we measure the resulting change in the output distribution. For example, we might replace an object in an image with a different object or mask a word in the text prompt. By comparing the original output with the counterfactual output, we can compute the average causal effect of that input feature on the prediction. Second, we attribute this effect to intermediate latent variables by performing activation patching: we swap the activation of a particular layer in the counterfactual run with the corresponding activation from the original run, and observe whether the output reverts to the original. This technique, known as causal tracing, reveals which layers play a critical role in encoding the visual information that leads to a correct or hallucinated response.

In multimodal settings, the causal graph becomes more complex because visual and textual pathways interact through cross-attention. A hallucination about an object that does not exist in the image may originate from the language model's overreliance on the text prompt, or from a misinterpretation of visual features that are present. Causal path tracing can disentangle these contributions by separately intervening on the visual pathway, the textual pathway, and the cross-modal pathway. For instance, by corrupting only the visual embeddings while keeping the text embeddings intact, we can determine whether a given hallucination is driven by visual misperception or by a language prior. Similarly, intervening on the cross-attention weights can reveal whether the model is attending to irrelevant visual regions.

The granularity of causal tracing can be adjusted based on computational constraints. At the coarsest level, we trace paths at the layer level, treating each transformer block as a node. At a finer granularity, we consider individual attention heads or even individual neurons. Finer granularity provides more precise localization of hallucination sources but incurs higher computational cost because it requires many more intervention experiments. The trade-off between granularity and efficiency is a central design consideration in the CausalRoute framework. For real-time or near-real-time applications, such as autonomous driving systems that must generate object descriptions with low latency, coarse-grained tracing may be sufficient to identify and suppress common hallucination patterns. For offline auditing and model debugging, finer-grained tracing can offer deeper insights.

4. Hallucination Suppression via Causal Intervention

Once causal paths responsible for hallucinations are identified, the next step is to suppress those paths through targeted interventions. In the CausalRoute framework, intervention is performed at the latent representation level rather than at the input or output level. This is motivated by the observation that hallucinations often arise from spurious correlations embedded in the latent space, which are difficult to correct by modifying inputs or by post-hoc filtering. The goal of intervention is to shift the model's internal representations away from the spurious pathway and toward the correct causal pathway.

One approach is to apply additive or multiplicative interventions to the latent activations along the identified path. For example, if a particular attention head in the cross-attention module is found to be primarily responsible for attending to a region of the image that contains a confounding object, we can reduce the weight of that head's output during inference. This is akin to a form of controlled ablation, but done in a continuous manner so as not to completely disable the head's contribution to other, correct predictions. Alternatively, we can reweight the entire pathway by applying a scaling factor that suppresses the flow of information through the problematic nodes. These interventions are parameterized by a set of

learnable coefficients that are optimized on a validation set of examples where hallucination labels are known.

A more sophisticated intervention strategy uses counterfactual representations. Instead of directly modifying the activations, we replace the representations along the hallucinatory path with representations from a counterfactual scenario where the spurious correlation is absent. For instance, if the model hallucinates the presence of a person in an image because the visual features are similar to a training image containing a person, we can replace the visual representation with one that corresponds to the same scene without a person. This counterfactual representation can be generated by a separate model trained to perform counterfactual image generation, or by interpolating between the original representation and the representation of a clean image. The resulting output is then conditioned on the counterfactual representation, which should suppress the hallucination.

The effectiveness of causal intervention depends on the faithfulness of the structural causal model. If the causal graph incorrectly encodes the true relationships between modalities, the interventions may suppress correct predictions or fail to eliminate hallucinations. Therefore, the CausalRoute framework includes a validation step that measures the consistency of the causal model using do-calculus and backdoor adjustment criteria [14,15]. In practice, building an accurate causal graph for a large multimodal model is challenging because the model's internal representations are high-dimensional and the interactions are highly nonlinear. Approximation methods, such as using linear probes to estimate causal effects [16], are employed to make the problem tractable. Despite these approximations, empirical studies have shown that even coarse causal interventions can reduce hallucination rates by a significant margin while preserving overall model performance.

5. Architectural Design and System Trade-offs

Integrating CausalRoute into a multimodal foundation model requires careful architectural design. The framework can be implemented as an add-on module that sits between the transformer layers and the output decoder, or it can be baked into the training procedure. Each approach has distinct trade-offs. A post-hoc module that runs during inference is easier to deploy because it does not require retraining the base model. However, it adds latency and computational overhead, especially if fine-grained causal tracing is performed for every generated token. To reduce overhead, we can precompute causal graphs for common input types and store them in a cache, updating them only when the model parameters change. Alternatively, we can perform causal tracing on a subset of tokens—specifically those that the model itself is uncertain about, as indicated by high entropy in the output distribution.

If causal intervention is integrated into the training pipeline, the model can learn to avoid spurious pathways from the beginning. In this scenario, the training objective is augmented with a causal regularization term that penalizes the model for relying on latent representations that are causally unrelated to the correct output. This approach is inspired by invariant risk minimization and causal representation learning [18]. The advantage is that the model internalizes the causal structure, reducing the need for online intervention during inference. The disadvantage is that training becomes more computationally intensive and may require additional data with causal annotations. For many real-world applications, the post-hoc inference module is more practical because it can be applied to existing pretrained models without the expense of retraining.

Another important architectural consideration is the granularity of the latent representations used for causal tracing. Most multimodal transformers use a bottleneck architecture where multimodal features are fused through cross-attention. A natural choice is to trace causal paths through the cross-attention outputs, because these are the points where visual and textual information are integrated. However, the cross-attention layer outputs are high-dimensional, and reducing them to a manageable number of causal nodes requires aggregation or dimensionality reduction. One approach is to use clustering methods to group similar attention patterns and treat each cluster as a node. Another is to use attention rollout, which aggregates attention weights across layers to create a coarse spatial map of where the model is focusing. The choice of aggregation method affects the resolution of causal tracing and the accuracy of the resulting interventions.

System deployment in a cloud or edge environment introduces further trade-offs. For latency-sensitive applications such as real-time captioning for autonomous vehicles, causal intervention must complete within milliseconds. This necessitates using lightweight causal models—for example, linear approximations of the causal graph—and performing interventions only on the most critical paths as identified by a separate, lightweight classifier trained to detect imminent hallucination. In batch processing scenarios, such as annotating medical images offline, deeper causal analysis can be performed with greater computational resources. The CausalRoute framework is designed with a modular interface that allows swapping between different tracing and intervention strategies depending on the deployment context.

6. Deployment, Infrastructure, and Governance

Deploying CausalRoute in production systems requires careful attention to infrastructure and governance. The causal tracing module generates detailed logs of which causal paths were intervened upon and the resulting changes in output. These logs serve as audit trails that can be used to verify the system's behavior and to demonstrate compliance with regulatory requirements. For example, under the European Union's AI Act, high-risk AI systems must provide transparency about how decisions are made and allow for human oversight. CausalRoute's interpretable outputs make it easier to explain why a model generated a particular caption or answer, and to show that hallucination suppression was applied in a principled manner.

Governance of causal interventions also involves monitoring for unintended side effects. A causal intervention that suppresses a spurious pathway might inadvertently weaken the model's ability to correctly represent rare but genuine correlations. For instance, suppressing a visual pathway that is associated with a certain texture might cause the model to misclassify objects that legitimately have that texture. To mitigate this risk, we propose a governance framework that includes continuous validation on held-out test sets, with periodic re-evaluation of the causal graph. If the model's performance on a set of canonical tasks drops below a threshold, the intervention coefficients are rolled back or recalibrated. This is analogous to a feedback control loop, where the system monitors its own behavior and adjusts accordingly.

Infrastructure requirements for CausalRoute include sufficient GPU memory to store intermediate activations during causal tracing, as well as high-speed interconnects to support the intervention operations. In cloud data centers, this can be accommodated by using dedicated inference servers that are optimized for large model inference. For edge deployment, where memory and compute are constrained, lightweight versions of CausalRoute can be

distilled from the full model. These distilled versions are trained to mimic the behavior of the intervened model on a representative set of inputs, thus avoiding the need to run full causal tracing on the device. The trade-off is that the distilled model may not generalize as well to unseen hallucination patterns.

7. Robustness, Fairness, and Sustainability

Hallucination suppression via causal intervention must be evaluated not only for effectiveness but also for its impact on model robustness and fairness. A potential concern is that causal intervention might disproportionately affect outputs for certain demographic groups or rare classes. For example, if the causal graph is derived from a training dataset that underrepresents certain visual concepts, the intervention might incorrectly suppress pathways that are actually correct for those concepts. This could lead to biased deletion of valid information, exacerbating existing fairness issues. To address this, the CausalRoute framework incorporates fairness audits that measure the model's performance across subpopulations before and after intervention. If disparate impact is detected, the intervention coefficients are adjusted to reduce bias.

Another aspect of robustness is the stability of causal interventions under distribution shift. A causal graph learned from one data distribution may not be valid for another. For instance, a model trained on natural images may hallucinate less after intervention on those images, but when deployed in a medical imaging context with different data characteristics, the same intervention may be ineffective or harmful. The framework addresses this by using domain adaptation techniques that update the causal graph based on a small amount of labeled data from the target domain. This requires the deployment infrastructure to support lightweight retraining of the causal module.

Sustainability is a growing concern in large-scale AI systems. Causal tracing and intervention involve additional forward passes through the model, increasing energy consumption. To minimize environmental impact, the CausalRoute framework employs early-exit strategies: for inputs where the model's confidence is high, causal tracing is skipped entirely. For uncertain inputs, only the most likely hallucination pathways are traced, using a greedy selection algorithm. Preliminary experiments suggest that this reduces the energy overhead by up to 60 percent while maintaining hallucination suppression performance. Additionally, the use of distilled edge models reduces the overall compute footprint. These measures align with broader efforts to make AI more sustainable.

8. Policy Implications and Future Directions

The development of methods like CausalRoute has significant policy implications. Regulators are increasingly calling for AI systems to be explainable and auditable, particularly in high-risk domains such as healthcare, criminal justice, and employment. Causal inference provides a rigorous framework for explainability, as it can answer counterfactual questions: "What would the model have output if the visual input had been different?" This capability is essential for accountability, as it allows stakeholders to understand the chain of reasoning leading to a decision. CausalRoute's ability to produce audit logs of interventions further supports compliance with transparency requirements.

Future directions include extending causal tracing to fine-grained temporal reasoning in video models and to multi-step reasoning chains in scientific applications. Another promising avenue is the integration of causal intervention into reinforcement learning from human feedback, where human annotators could flag hallucinated outputs and a causal model could

be used to propagate those corrections to similar inputs. The combination of causal inference with continual learning could enable models to adapt their causal graphs over time as they encounter new data distributions.

One limitation of the current framework is its reliance on a manually defined causal graph structure. Future work could explore automated discovery of causal graphs from the model's own representations, using techniques such as structural equation modeling or differentiable causal discovery. This would make the framework fully autonomous and more scalable. Finally, as multimodal foundation models become more widely deployed in safety-critical systems, the CausalRoute approach represents an important step toward ensuring that these systems are not only powerful but also trustworthy.

9. Conclusion

This paper presented CausalRoute, a framework for causal path tracing and hallucination suppression in multimodal foundation models. By leveraging structural causal models, the framework identifies the latent pathways through which spurious correlations propagate and intervenes on those pathways to reduce hallucinated content. We discussed the architectural trade-offs involved in implementing such a system, the infrastructure and governance requirements for deployment, and the implications for robustness, fairness, and sustainability. Causal inference offers a principled and interpretable approach to improving the reliability of multimodal AI systems. As these systems continue to enter high-stakes applications, the need for causal grounding will only grow. CausalRoute provides a practical foundation for building more transparent, accountable, and trustworthy AI.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4171–4186.

3. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. International Conference on Machine Learning, 8748–8763.

4. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

5. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10684–10695.

6. Li, J., Li, D., Xiong, C., & Hoi, S. C. H. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. International Conference on Machine Learning, 12888–12900.

7. Shi, C., Li, S., Lu, W., Wu, W., Wang, C., Cheng, Z., ... & Chua, T. S. (2026). TraceRouter: Robust Safety for Large Foundation Models via Path-Level Intervention. arXiv preprint arXiv:2601.21900.

8. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

9. Li, Y., Du, Y., & Liang, P. (2020). A simple and effective approach for hallucination detection in image captioning. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 869–879.

10. Rohrbach, A., Rohrbach, M., & Schiele, B. (2015). The long-tail of hallucination in visual captioning: A new benchmark and analysis. Proceedings of the IEEE International Conference on Computer Vision, 2201–2209.

11. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. W. (2020). REALM: Retrieval augmented language model pre-training. International Conference on Machine Learning, 3929–3938.

12. Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. International Conference on Learning Representations.

13. Su, Y., Lan, T., Wang, Y., Yogatama, D., Kong, L., & Collier, N. (2022). A contrastive framework for neural text generation. Advances in Neural Information Processing Systems, 35, 21548–21562.

14. Pearl, J. (2009). Causality: models, reasoning, and inference (2nd ed.). Cambridge University Press.

15. Peters, J., Janzing, D., & Schölkopf, B. (2017). Elements of causal inference: foundations and learning algorithms. MIT Press.

16. Vig, J., Gehrmann, S., Deng, Y., Sap, M., Wiegreffe, S., & Belinkov, Y. (2020). Causal mediation analysis for interpreting neural networks: A case study on transformer models. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 7325–7341.

17. Geiger, A., Wu, Z., Lu, H., Rozner, J., Kreiss, E., Icard, T., ... & Potts, C. (2021). Inducing causal structures for interpretable neural networks. International Conference on Machine Learning, 3675–3685.

18. Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. arXiv preprint arXiv:1907.02893.

19. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

20. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, 618–626.