Towards Vision-Language Mechanistic Interpretability: A Causal Tracing Tool for BLIP

Vedant Palit* IIT Kharagpur
Rohan Pandey* Reworkd.ai
Aryaman Arora Georgetown University
Paul Pu Liang Carnegie Mellon University

Abstract

Mechanistic interpretability seeks to understand the neural mechanisms that enable specific behaviors in Large Language Models (LLMs) by leveraging causality-based methods. While these approaches have identified neural circuits that copy spans of text, capture factual knowledge, and more, they remain unusable for multimodal models since adapting these tools to the vision-language domain requires considerable architectural changes. In this work, we adapt a unimodal causal tracing tool to BLIP to enable the study of the neural mechanisms underlying image-conditioned text generation. We demonstrate our approach on a visual question answering dataset, highlighting the causal relevance of later layer representations for all tokens. Furthermore, we release our BLIP causal tracing tool as open source to enable further experimentation in vision-language mechanistic interpretability by the community. Our code is available at this URL.

1. Introduction

Mechanistic interpretability [30] analyzes neural networks with the goal of reverse engineering the algorithms a network implicitly learns in their parameters. This allows for finer-grained control over a model's knowledge [27, 28, 17] and behavior [23]. In particular, causal mediation analysis (CMA) [34] is a popular mechanistic interpretability method that studies the effect of introducing a mediator on the outcome of a system. However, CMA has so far been implemented only for the unimodal language domain [27], limiting our understanding to this narrow class of models [5].

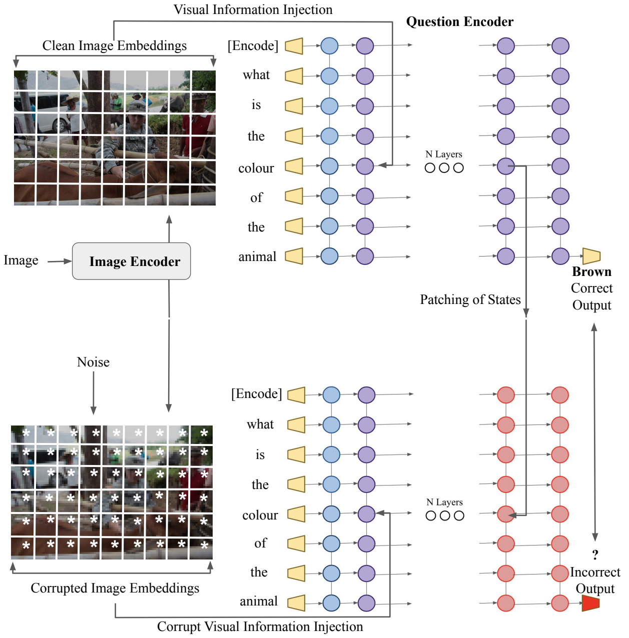


Figure 1: Causal intervention to measure state's relevance: Above, an image of a cow is encoded, cross-attends with the question encoding, and results in the correct answer "brown". Below, the same image encoding is corrupted, cross-attends with the question encoding, and results in an incorrect answer. An intermediate state is patched from the clean to the corrupted run to observe the state's effect on the answer probabilities.

In recent years, multimodal models have rapidly grown in relevance as vision-language transformers have enabled strong performance on image-text retrieval, image captioning, and visual question answering (VQA) tasks [25]. Considering the powerful effects of visual stimulus on semantic representations in humans [21], it is important to understand how similar processes occur in vision-language models. Take as an example the vision-language transformer BLIP [22], which consists of an image encoder cross-attending with a text encoder, jointly conditioning a text decoder (Fig. 2). In this work, we seek a deeper understanding of how BLIP performs VQA by adapting CMA to the vision-language setting.

*Corresponding authors: vedantpalit@kgpian.iitkgp.ac.in, rohan@reworkd.ai

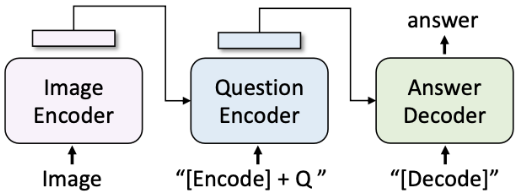
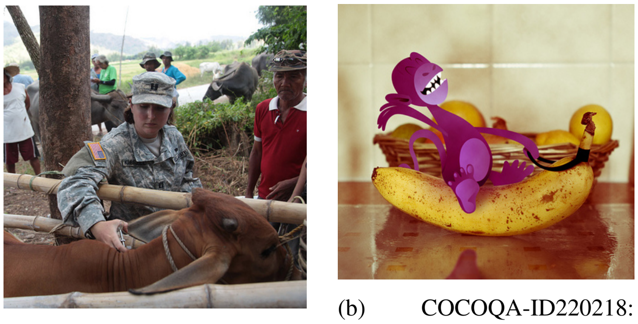


Figure 2: The BLIP-for-VQA [22] architecture: embedding for an image patch is fed into the question encoder alongside question tokens to generate image-conditioned question embeddings through cross-attention, which are finally input to the answer decoder for answer generation.

Task	Accuracy
Color Identification	**80.23%**
Location Identification	26.30%
Object Counting	3.27%

Table 1: BLIP Performance on COCO-QA Task Categories



(a) COCOQA-ID458864: What is the color of the animal?

(b) COCOQA-ID220218: What is the color of the character?

Figure 3: Two example images from COCO-QA and their accompanying questions.

2. Related Work

Pearl [34] introduces causal mediation analysis by measuring the change in a response variable following an intervention, taking into consideration the effects of intermediaries or mediators. Vig et al. [40] applies this analysis to language models of the GPT-2 family to study how grammatical gender bias is mediated by the different components inside a model. They argue that probing representations [1, 14, 9] for information does not tell us [2, 37] whether the model actually uses this information, and causal approaches to interpretability are a better approach.

Meanwhile, researchers in mechanistic interpretability have developed a variety of techniques to better understand neurons and mechanisms inside neural networks (particularly unimodal language models), building on earlier work on identifying circuits in vision models [31]. This includes applying linear algebra to understand interactions between modules inside the transformer architecture [11, 32], studying the training dynamics of transformer models, often on simple tasks [10, 29, 3, 7, 19], intervening on model-internal activations to identify causal relationships between model components [12, 41, 13, 42, 8, 15], and attempting to map neuron features to human-interpretable concepts [43, 16, 4].

Meng et al. [27] also base their causal intervention methods on the previous works by corrupting token embedding inputs to a language model (GPT-2 XL, GPT-NeoX) to measure causal relevance of states for capturing factual knowledge. The corruption in the input is produced by introducing noise into a sentence's subject tokens. Following this, the models are observed in three different runs—a clean input run, a corrupted input run, and an intervention involving patching of the layer outputs from a clean run of the same sentence input to the corresponding layer outputs of a corrupted run. Our implementation follows this work most closely.

On the multimodal side of interpretability, there have been thorough analyses of vision-language transformers leveraging probing approaches [6, 36], though these face the same epistemic issues as those in the unimodal setting [2]. Another line of work explores unimodal interactions present in a multimodal model and proposes methods to understand the nature & degree of these interactions [39, 24]. Joshi et al. [18] present a comprehensive survey of interpretability in multimodal machine learning until early 2021. Finally, Kervadec et al. [20] present some interpretability experiments on transformers trained for VQA, specifically concerned with their reasoning ability.

3. Method

We adapt the causal intervention method from Meng et al. [27] to investigate visual question answering (COCO-QA) in a vision-language model (BLIP).

3.1. Causal Tracing for BLIP

As input, BLIP takes a pre-processed image and question tokens, returning a single-word answer as output. We corrupt the image embeddings before they are fed into the question encoder, resulting in an incorrect output. Following this, we try and 'make the answer correct again' by patching individual intermediate states (token embeddings at a layer) of a clean run into the corrupted run. The states that result in the greatest answer improvement are considered causally relevant.

Corruption and Patching The second image embedding of the batch is corrupted by adding noise to all the 577 patch embeddings of the image, resulting in pairs of clean and corrupted embeddings (E, E^*). For each image, we sample a single instance of noise $\epsilon \sim \mathcal{N}(1, \nu)$, where ν is an adjustable hyperparameter (standard deviation of noise), and multiply ϵ to the embedding for each patch. This corrupted image encoding is then passed into the question encoder alongside question input tokens for patching.

To perform the causal intervention, the output of each individual state (layer L, token T) of the E^* question encoder run is overwritten with the corresponding state from the clean image embedding run E (see Fig. 1). Finally, we measure the resulting effect on output logits. This intervention process is also replicated for the answer decoder block.

Metrics Given the question text embedding Q and the image embedding pair (E, E^*), to measure the effect of our causal intervention, we compare the correct answer's (A) probability between the corrupted run $p(\mathrm{A} \mid E^*, \mathrm{Q})$ and the restored run (where we patch from the clean run into the corrupted run at layer L and token position T), and normalize across the difference between the clean and corrupted run probabilities:

$$\Gamma_{L,T} = \frac{p(\mathrm{A} \mid \mathrm{patch}_{L,T}(E, E^*), \mathrm{Q}) - p(\mathrm{A} \mid E^*, \mathrm{Q})}{p(\mathrm{A} \mid E, \mathrm{Q}) - p(\mathrm{A} \mid E^*, \mathrm{Q})} \quad (1)$$

We expect $\Gamma_{L,T}$ to be in the range $[0, 1]$, where 0 represents no improvement from complete corruption and 1 represents perfect recovery of the original answer probability.

We may then plot $\Gamma_{L,T}$ for all (L, T) pairs to observe the causal relevance of that state on producing the correct answer. The darker shades of the heatmaps in Fig. 4 represent high causal relevance $\Gamma_{L,T}$. We can also compute an average probability difference as a function of the noise factor:

$$\Gamma(\nu) = \frac{1}{|L| \cdot |T|} \sum_{l \in L} \sum_{t \in T} \Gamma_{l,t}(E^* = \nu E) \quad (2)$$

We plot this function in Fig. 5, illustrating how the average difference in answer probabilities varies depending on the strength of the image embedding's corruption noise.

3.2. COCO-QA Dataset

VQA is an open-ended answer generation task which requires the model to predict an answer given an image and associated question input. We utilize this task as a simple testbed for causal tracing vision-language models. The dataset used we use is COCO-QA [35] consisting of 123,287 images, 78,736 train and 38,948 test questions. This was sourced from MSCOCO [26]. The COCO-QA dataset contains one-word answers to questions belonging to four categories: object identification, object counting, colour identification, and location identification.

We divided the training subset of COCO-QA into three splits pertaining to each of the three categories: colour, location identification, and counting. Following this division, BLIP's zero-shot performance was assessed on each of the datasets individually, results of which are shown in Table 1.

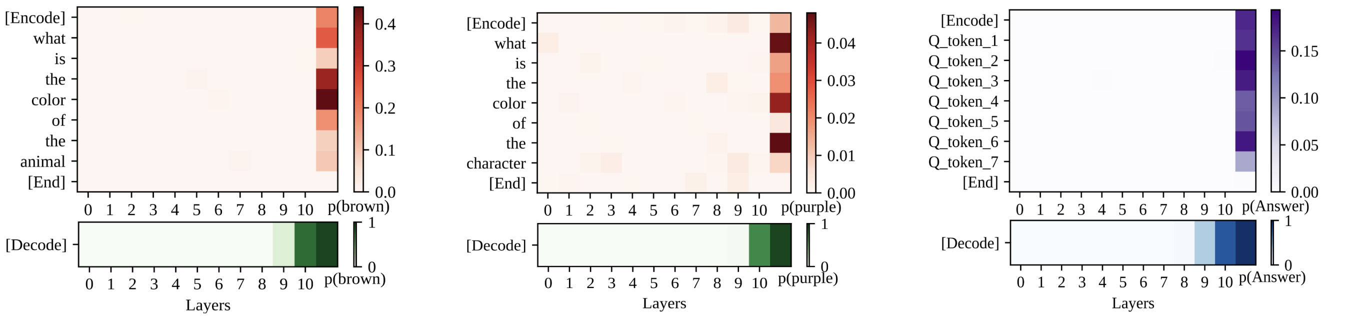
The accuracy percentages demonstrate that BLIP's pretrained VQA model performs best in the color identification task. Further analysis showed that BLIP tends to output number of objects in an image using digits rather than natural language, which causes a low accuracy score on textual answers. Similarly, it also differs in answer structuring in the location identification task. Thus, we utilize the color identification data split of COCO-QA for causal tracing, since we want to understand mechanisms behind a behavior that a model is highly performant at.

4. Results

In order to understand the correlation between the amount of noise injected into the image embeddings with $\Gamma(\nu)$, we first plotted the effects of adjusting the noise factor ν in the range $[0.1, 30]$, averaged over 200 samples from the dataset with 10 runs for each of the samples (see Fig. 5). We do not measure Γ when ν is 0, since we would be patching from clean runs into clean runs, so $\Gamma(\nu) = 1$. A decaying curve is observed as the ν value increases from 0.1 to 30, with very little variation in $\Gamma(\nu)$ at extremely large values and negative values for a few values of ν. Keeping both the curves in mind, we refrain from injecting too little noise that patching becomes trivial or too much noise where restoration becomes impossible, hence choosing ν as 5.

The heatmaps in Fig. 4a and Fig. 4b demonstrate the causal effects in the question encoder and answer decoder for two examples from the dataset shown in Fig. 3, averaged across 10 runs. Fig. 4c demonstrates the average effects across 200 samples from the COCO-QA dataset. The encoder and decoder layers are indexed from 0–11, and input question tokens are plotted vertically.

It is clear from the figures that in the question encoder, only the final layer (11) for all tokens plays a significant role in affecting the output to a higher degree than any preceding layers or tokens. In the case of the answer decoder, the final layers (9 to 11) play the most apparent role in the final output of the model. These results show that BLIP does not benefit from restored access to the correct image embeddings until the final few layers. This may mean that the vision modality is not relevant to model computations until the final layer, i.e. vision and language are processed independently in the intermediate layers. On the other hand, it may also mean that the final layers override preceding layers, which may still be weakly causally relevant to the

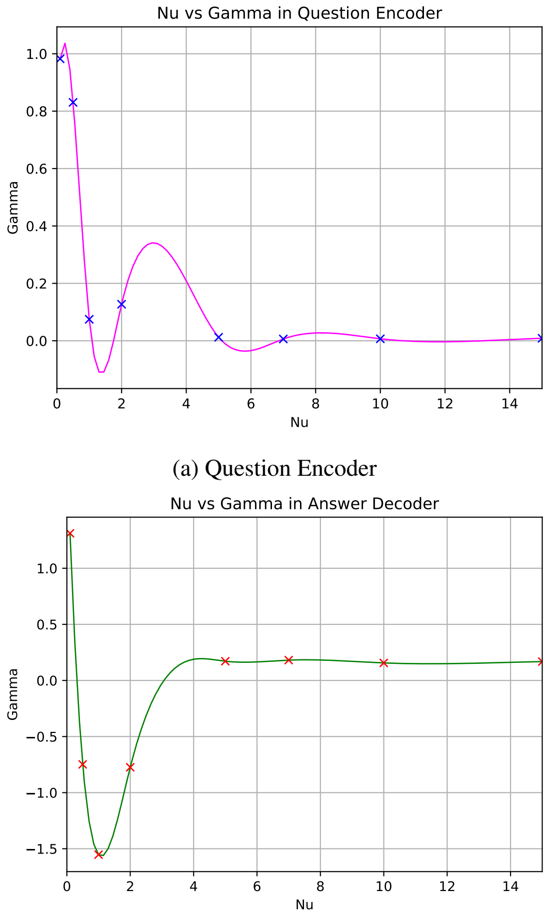


(a) COCOQA-ID458864

(b) COCOQA-ID220218

(c) Average over 200 samples

Figure 4: Probability $\Gamma_{L,T}$ of the correct answer after performing causal interventions at specific layers on specific tokens in the question encoder (above) and answer decoder (below). Most of the causal relevance is concentrated in the final layers of the encoder as well as decoder blocks.



(a) Question Encoder

(b) Answer Decoder

Figure 5: Effect of adjusting the noise factor ν on the answer probability difference Γ (averaged across all L, T patches) for different components of the BLIP model.

model output.

5. Conclusion

We introduced the first causal tracing tool for a vision-language model and studied how model performance is localized in BLIP on a subset of the visual question answering task with the COCO-QA dataset. Previous work on interpretability of vision-language models has not focused on identifying causal mechanisms, so we hope that this work invigorates research in this area. Towards this end, we fully open source our code and will soon release a visualizer as well as adaptations to other vision-language models.

Many aspects of the causal tracing methodology are still not fully understood. For example, since the role of the noise factor ν is unclear, future work could study why different components of the model have different sensitivities to noise; for example, why is performance not monotonically reduced by increasing ν? Also, restoration of the clean image embedding at *multiple* points (instead of just one) may help us understand cross-module coordination within the model.

A bigger project is to identify larger mechanisms within vision-language models that can explain how the model performs specific tasks, as has been done in unimodal language models [29, 41]. This will help us understand how multi-modal models work and let us verify whether they perform tasks as expected, e.g. whether they learn good algorithms or poor shortcuts on compositional understanding benchmarks like Winoground [38, 33]. Overall, much work remains in this line of research and we look forward to using causal intervention methods for disentangling the mechanisms learned by vision-language models.

References

[1] Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained analysis of sentence embed-

dings using auxiliary prediction tasks, 2017.

[2] Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 2019.

[3] Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. *arXiv preprint arXiv:2306.00802*, 2023.

[4] Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. *OpenAI*, 2023.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[6] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 565–580. Springer, 2020.

[7] Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. *arXiv preprint arXiv:2302.03025*, 2023.

[8] Arthur Conmy, Augustine N Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *arXiv preprint arXiv:2304.14997*, 2023.

[9] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[10] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

[11] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

[12] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586. Curran Associates, Inc., 2021.

[13] Atticus Geiger, Chris Potts, and Thomas Icard. Causal abstraction for faithful model interpretation. *arXiv preprint arXiv:2301.04709*, 2023.

[14] Mario Giulianelli, Jacqueline Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information, 2021.

[15] Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

[16] Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*, 2023.

[17] Evan Hernandez, Belinda Z Li, and Jacob Andreas. Measuring and manipulating knowledge representations in language models. *arXiv preprint arXiv:2304.00740*, 2023.

[18] Gargi Joshi, Rahee Walambe, and Ketan Kotecha. A review on explainability in multimodal deep neural nets. *IEEE Access*, 9:59800–59821, 2021.

[19] Jeevesh Juneja, Rachit Bansal, Kyunghyun Cho, João Sedoc, and Naomi Saphra. Linear connectivity reveals generalization strategies. *arXiv preprint arXiv:2205.12411*, 2022.

[20] Corentin Kervadec, Theo Jaunet, Grigory Antipov, Moez Baccouche, Romain Vuillemot, and Christian Wolf. How transferable are reasoning patterns in vqa? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4207–4216, 2021.

[21] George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago press, 2008.

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[23] Maximilian Li, Xander Davies, and Max Nadeau. Circuit breaking: Removing model behaviors with targeted ablation. In *DeployableGenerativeAI*, 2023.

[24] Paul Pu Liang, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zihao Deng, Xingbo Wang, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multiviz: Towards visualizing and understanding multimodal models. In *The Eleventh International Conference on Learning Representations*, 2022.

[25] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*, 2022.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[27] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc., 2022.

[28] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan

Belinkov, and David Bau. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2022.

[29] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023.

[30] Chris Olah. Mechanistic interpretability, variables, and the importance of interpretable bases. *Transformer Circuits Thread*, 2022.

[31] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

[32] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

[33] Rohan Pandey. Semantic composition in visually grounded language models. *arXiv preprint arXiv:2305.16328*, 2023.

[34] Judea Pearl. Direct and indirect effects. In *Probabilistic and causal inference: the works of Judea Pearl*, pages 373–392. 2022.

[35] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28, 2015.

[36] Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. Are vision-language transformers learning multimodal representations? a probing perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11248–11257, 2022.

[37] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline, 2019.

[38] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.

[39] Yao-Hung Hubert Tsai, Martin Q Ma, Muqiao Yang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multimodal routing: Improving local and global interpretability of multimodal language analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2020, page 1823. NIH Public Access, 2020.

[40] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

[41] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 Small. In *The Eleventh International Conference on Learning Representations*, 2023.

[42] Zhengxuan Wu, Atticus Geiger, Christopher Potts, and Noah D. Goodman. Interpretability at scale: Identifying causal mechanisms in Alpaca. *arXiv preprint arXiv:2305.08809*, 2023.

[43] Roland S Zimmermann, Thomas Klein, and Wieland Brendel. Scale alone does not improve mechanistic interpretability in vision models. *arXiv preprint arXiv:2307.05471*, 2023.