LVLM-Intrepret: An Interpretability Tool for Large Vision-Language Models

Gabriela Ben Melech Stan[1*] Estelle Aflalo[1*] Raanan Yehezkel Rohekar[1*] Anahita Bhiwandiwalla[1*]
Shao-Yen Tseng[1*] Matthew Lyle Olson[1*] Yaniv Gurwicz[1*] Chenfei Wu[2] Nan Duan[2] Vasudev Lal[1]
[1]Intel Labs [2]Microsoft Research Asia
https://intellabs.github.io/multimodal_cognitive_ai/lvlm_interpret/

Abstract

In the rapidly evolving landscape of artificial intelligence, multi-modal large language models are emerging as a significant area of interest. These models, which combine various forms of data input, are becoming increasingly popular. However, understanding their internal mechanisms remains a complex task. Numerous advancements have been made in the field of explainability tools and mechanisms, yet there is still much to explore. In this work, we present a novel interactive application aimed towards understanding the internal mechanisms of large vision-language models. Our interface is designed to enhance the interpretability of the image patches, which are instrumental in generating an answer, and assess the efficacy of the language model in grounding its output in the image. With our application, a user can systematically investigate the model and uncover system limitations, paving the way for enhancements in system capabilities. Finally, we present a case study of how our application can aid in understanding failure mechanisms in a popular large multi-modal model: LLaVA.

1. Introduction

Recently, large language models (LLM), such as those in the families of GPT [24] and LLaMA [39, 40], have demonstrated astounding understanding and reasoning capabilities, as well as the ability to generate output that adheres to human instructions. Building on this ability, many work, such as GPT-4V [25], Qwen-VL [4], Gemini [37], and LLaVA [14], have introduced visual understanding to LLMs. Through the addition of a vision encoder followed by finetuning on multimodal instruction-following data, these prior work have demonstrated large vision-language models (LVLM) that are able to follow human instructions to complete both textual and visual tasks with great aptitude.

LLMs are rapidly surpassing humans in many tasks such as summarization, translation, general question answering, and even creative writing. However, they are still very prone to hallucination, *i.e.* the fabrication of untrue information [12, 45]. This phenomenon of hallucination is also seen in LVLMs and may even include additional dimensions stemming from the visual modality [42, 50]. With the introduction of LVLMs and their massively increased number of parameters, interpreting and explaining model outputs to mitigate hallucination is becoming an ever rising challenge. In light of the need to understand the reasoning behind model responses, we present an intrepretability tool for large vision-language models: LVLM-Intrepret. The proposed application adapts multiple interpretability methods to large vision-language models for interactive analysis. These methods include raw attention, relevancy maps, and causal interpretation. LVLM-Interpret is applicable to any LVLM with a transformer-based LLM front-end. We further demonstrate how we can gain insight on the inner workings of LVLMs using our application.

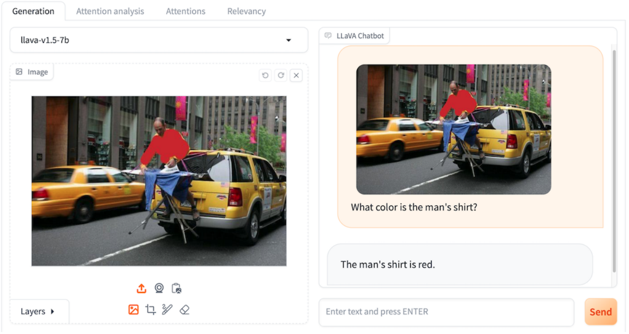
The main contributions of this paper are:

- We propose an interactive tool for interpreting the inner attention mechanisms of large vision-language models
- We present a case study that sheds light on a possible cause behind certain failure cases in LVLMs
- Through a study on causal explanations, we postulate that large vision-language models (such as LLaVA [14]) implicitly learn to represent causal structure

2. Related Work

The advancements in deep learning models has been preceded by novel interpretability and explainability tools to better understand the internal workings of these models. Earlier works [43, 48, 49] demonstrated the use of explanatory graphs, decision trees, histograms, respectively to analyze machine learning models. As Transformer [41] based architectures gained popularity in the field, various approaches such as [7] proposed computing relevancy scores across the layers of the model, [31] generalized the attention from low-level input features to high-level concepts to ensure interpretability within a specific domain, while [26]

[*] Main authors



What color is the man's shirt?

The man's shirt is red.

Figure 1. Main interface of LVLM-Interpret. Users can issue multimodal queries using a chatbot interface. Basic image-editing feature allows for model probing.

presented a novel interpretability-aware redundancy reduction transformer framework.

2.1. Interpretability of Vision Models

Studying the intepretability of Vision Transformers (ViT) has gained popularity with task-specific analysis like image captioning [8, 10, 35], object detection [3, 9, 44], image recognition [19, 46]. Recently, there has been an increased demand for interpretability analysis for the medical domain in applications such as pathology [13, 22], retinal image classification [11, 27] and COVID-19 analysis [20, 34] among others. A novel vision transformer was presented in [28] with a training procedure which has an interpretability-aware training objective. [21] proposed a method to use the activations of ViT's hidden layers to predict the relevant parts of the input that contribute to its final predictions, while [17] introduced quantification indicators to measure the impact of patch interactions to effectively exploit responsive fields of patches in ViT.

2.2. Interpretability of Multimodal Models

Multimodal models have proliferated various domains across healthcare, multimedia, industrial applications among others. There has been a rise in independent interpretability studies of such multimodal systems [2, 6, 15, 16, 29, 36]. For the medical domain where the reasoning behind decisions, high stakes involved are of utmost importance. [18] demonstrated the use of an interpretability method based on attention gradients to guide the transformer training in a more optimal direction, while [5] presented an interpretable fusion of structural MRI and functional MRI modalities to enhance the accuracy of schizophrenia.

Our hope with this proposed interpretability tool is not to replace domain-specific solutions, but to complement them, improving existing and future large vision language models and further strengthen the confidence in the predictions and behavior of these models.

3. Interface and Interpretability Functions

LVLM-Intrepret was developed using Gradio [1] and follows a standard layout for multimodal chat. Figure 1 shows an example of the user interface. In the UI, a user is able to upload an image and issue multimodal queries to the LVLM. An added editing feature allows for basic modification of the input image to probe the model with adversarial variations. As the LVLM model generates a response, the attention weights of the model are stored internally and are later presented to the user for visualization. The application also constructs relevancy maps and causal graphs relating to the output of the model. Once a response is returned, the user is able to utilize these results as a way to interpret the model output. The following sections describe each of these interpretability functions.

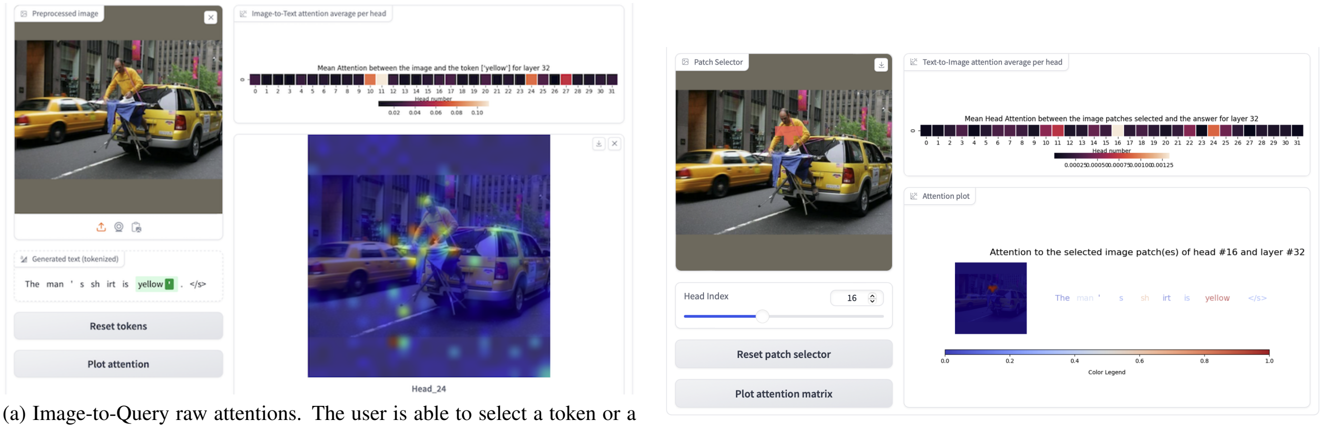
3.1. Layer Attentions

Following work such as VL-Interpret [2], LVLM-Interpret also allows for interactive visualization of raw attentions. More specifically, our application allows users to investigate the interactions among tokens from each modality. Heatmaps that show average attentions between image tokens and query tokens as well as answer tokens enables the user to better understand the global behavior of raw attentions. Figure 2 shows how a user can visualize raw attentions for a specific head and layer. As shown in Figure 2a, the user can select tokens from the generated response and visualize the average attentions between image patches and the selected tokens to obtain insight on how the model attends to the image when generating each token. Conversely, Figure 2b shows how a user can select image patches and visualize the degree to which each output tokens attends to that specific location.

3.2. Relevancy Map

Relevancy maps [6, 7] aim at interpreting the decision-making process of transformers. These maps are designed to enhance interpretability by illustrating how different components of an input, whether text or image, are relevant to the model's generated output, overcoming some limitations of traditional attention visualization techniques. The method assigns a local relevancy scores to each element in the input based on their contribution to the output decision. We refer the reader to [6] for more details on the approach.

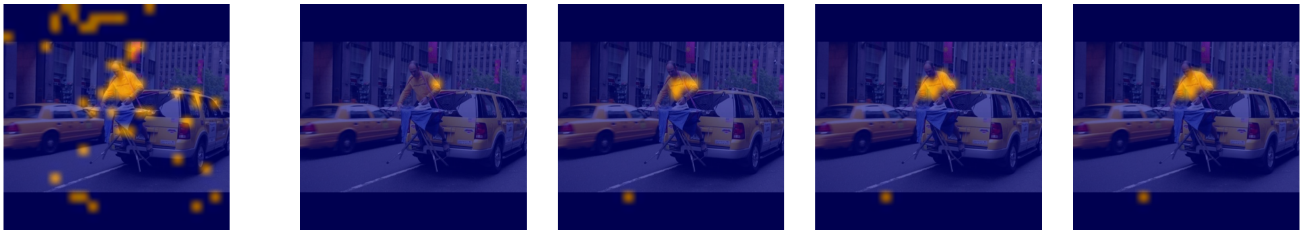
We adapted the calculation of relevancy maps to LVLMs such as LLaVA. Relevancy scores are backward propagated through the LLM as well as vision transformer commonly used as the vision encoder. For image analysis, the relevancy scores corresponding to image patches are reshaped into a grid that matches the layout of the original image. This grid forms the basis of the relevancy map. The relevancy map is then upscaled to the original image size using



(a) Image-to-Query raw attentions. The user is able to select a token or a group of tokens to visualize the attention values of image to text output for each head and layer.

(b) Query-to-Image raw attentions. The user is able to select image patches to visualize attention values going into answer tokens.

Figure 2. Visualization of crossmodal attentions



(a) Raw high-Attention

(b) Search distance = 1

(c) Search distance = 2

(d) Search distance = 3

(e) Search distance = 4

Figure 3. Causality-based explanation for the token ‘`yellow`’ in the generated answer ‘`The man’s shirt is yellow`’ at head 24. (a) Top 50 image-tokens having the highest raw attention values. Each serves as a graph node. (b-e) Image tokens from the explanation set identified by the CLEANN method, at different search distances on the learned causal graph. Tokens are marked with yellow blobs.

bilinear interpolation, providing a visualization of the regions of the image most relevant to each generated token.

Relevancy maps can aid in model debugging, ensuring fairness, and providing explanations for inaccuracies by identifying the the most relevant parts of the input to the generated output, as demonstrated through a case study in Section 4.

3.3. Causal Interpretation

Recently, a causal interpretation of the attention mechanism in transformers was presented [33]. This interpretation leads to a method for deriving causal explanations from attention in neural networks (CLEANN). In this sense, if explanation tokens would have been masked in the input, the model would have generated a different output. Such explanations, which are a subset of the input tokens, are generally tangible and meaningful to humans. The method was previously demonstrated for a single modality, such as recommendation systems [23], and text sentiment classification [33]. Here, we enable examining if multi-modal LLMs, which are significantly larger, internally represent causal structures. Hence, in addition to providing causal explanation, we plot the causal graphs around the explained tokens, and allow the user to decrease or increase the explanation set size based on this graph. We refer the reader to [32, 33] for more details.

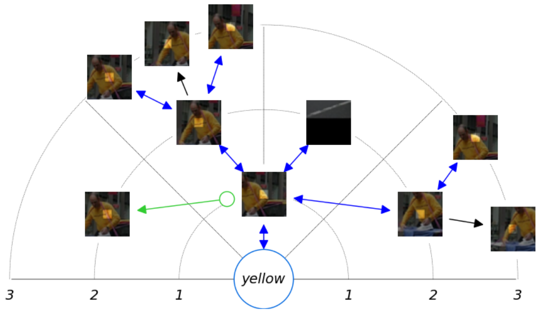
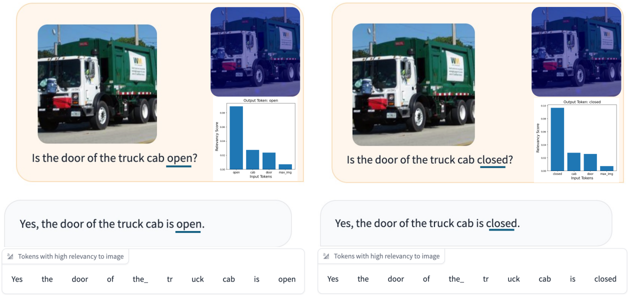


Figure 4. A tree constructed from the causal graph from which explanations for the token ’`yellow`’ are extracted. Arc radius indicates distance on the causal graph. Edges are color coded, bi-directed edges indicate a latent confounder, a circle edge-mark indicates that both a ‘tail’ and ’arrow’ are valid.

We employ CLEANN to explain large vision-language models by learning causal structures over input-output sequences of tokens. The presence of each token in this sequence is an event which is represented by a node in the



Is the door of the truck cab open?

Yes, the door of the truck cab is open.

Is the door of the truck cab closed?

Yes, the door of the truck cab is closed.

Output Token: open
Relevancy Score
0.00
0.02
0.04
0.06
0.08
open
cab
door
max_img
Input Tokens
Output Token: closed
0.10
closed
Tokens with high relevancy to image
Yes the door of the_ tr uck cab is open
Tokens with high relevancy to image
Yes the door of the_ tr uck cab is closed

Figure 5. Example where LLaVA seems to prioritize text input over image content. Presented with an unchanging image of a garbage truck, the model provides contradictory responses (‘yes, the door is open’ vs. ‘yes, the door is closed’) based on the query’s phrasing. Relevancy maps and bar plots for open and closed tokens demonstrate higher text relevance compared to image.

causal graph. Thus, the event of generating an output token is explained by the presence of a subset of input tokens.

Consider the following example. A sequence of image token is given as part of the prompt, and the text prompt is ‘What color is the man’s shirt?’. In response, the model generates ‘The man’s shirt is yellow.’. One may be interested in understanding which image tokens are responsible for the the token yellow. That is, identify the parts of the image such that if masked, will cause the model to output a different color. First, the top-k tokens having the highest attention values for yellow are assigned to nodes. Then, using the full attention matrix of the last (deepest) layer is used to learn a causal graph. This causal graph is a partial ancestral graph [30, 47], where a circle edge-mark indicates a non identifiable edge mark (head and tail are equally valid). From this graph, a tree rooted at node yellow is extracted (Figure 4) such that it includes all paths that potentially influence the root [33, Appendix B. Definition 2]. CLEANN searches for the minimal explaining set by gradually increasing the search distance in this tree (radius in Figure 4) from the explained node. An example is given in Figure 3. In Figure 3a, the 50 tokens having the highest attention values are marked. In Figure 3b–Figure 3e, tokens within different search distances from the explained token are marked.

This explanation approach solely relies on attention values in the last layer. While raw attention values describe pair-wise, marginal dependence relations, the causal-discovery algorithm in CLEANN identifies conditional independence relations. Thus, based only on the current trained weights, it can identify those tokens that if perturbed may change the generated token.

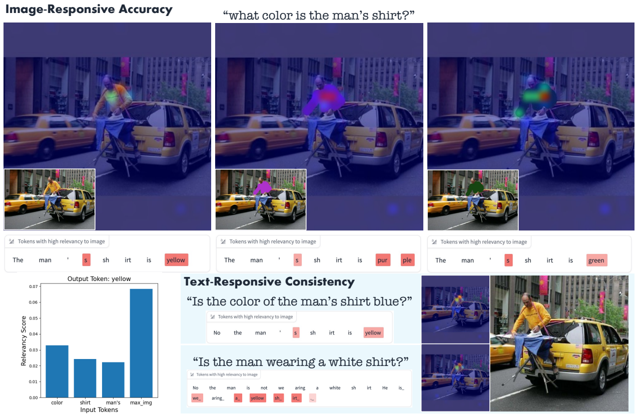


Image-Responsive Accuracy

“what color is the man’s shirt?”

Text-Responsive Consistency

“Is the color of the man’s shirt blue?”

“Is the man wearing a white shirt?”

Tokens with high relevancy to image
The man ' s sh irt is yellow
Tokens with high relevancy to image
The man ' s sh irt is pur ple
Tokens with high relevancy to image
The man ' s sh irt is green
Output Token: yellow
Relevancy Score
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
color
shirt
man's
max_img
Input Tokens
Tokens with high relevancy to image
No the man ' s sh irt is yellow
Tokens with high relevancy to image
No the man is not we aring a white sh irt He is_
we_ aring_ a_ yellow sh_ irt_ ._

Figure 6. Example where LLaVA demonstrates visual consistency with high relevancy scores for correct output tokens. With a constant query, ‘What color is the man’s shirt?’, manually altering the shirt’s color in the image (purple, green) changes the model’s answer which align with image changes. Relevancy scores highlight stronger connections to image than text tokens, illustrated by the image relevancy maps (upper row) and a bar plot comparison of relevancy scores. With a constant image, despite different question phrasings, the model demonstrates consistency in its answer, underscoring a strong relevancy of the yellow token to the visual input over the textual input variations.

4. Case Study

To demonstrate the functionalities of LVLM-Interpret, we analyze the LLaVA model on samples from the Multimodal Visual Patterns (MMVP) benchmark dataset [38]. MMVP focuses on identifying “CLIP-blind” images that are demonstrably hard for LVLMs to reason on. This dataset is of particular interest for our study since it highlights the challenges faced in answering relatively straightforward questions, often leading to incorrect responses.

We adapted the relevancy scores to examine the impact of both text and image tokens on the output generated by LLaVA-v1.5-7b. Given that the text and vision transformers responsible for generating the input embedding were kept frozen during LLaVA finetuning, our initial step involved calculating the relevancy scores for each generated output relative to the input features to LLaMA, focusing on the LLaMA self-attention layers. We observed instances where LLaVA mainly attends to the text tokens and less to the image tokens, indicated by lower relevancy scores to the image tokens relatively to relevancy scores to the input text tokens. In these cases, the model becomes more susceptible to manipulation, in some cases altering its responses based on the query with low regard to the image content. This phenomenon is exemplified by the truck scenario depicted in Figure 5. Conversely, when the generated outputs exhibit a greater relevance to image tokens than to input text, the accuracy of LLaVA appears to remain unaffected by how the question is phrased, as illustrated in Figure 6.

5. Conclusions and Future Directions

In this paper we presented LVLM-Interpret, an interactive tool for interpreting responses from large vision-language models. The tool offers a way to visualize how generated outputs relate to the input image through raw attention, relevancy maps, and causal interpretation. Through the many interpretability functions, users can explore the inner mechanisms of LVLMs and obtain insights on failure cases. The application can also reveal several potential paths for enhancing the performance of LVLMs. Future work can include consolidation of the multiple interpretability methods for a more comprehensive metric to explain the reasoning behind model responses.

References

[1] Gradio. https://www.gradio.app/. Accessed: 2024-03-13. 2

[2] Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21406–21415, 2022. 2

[3] Ji-Won Baek and Kyungyong Chung. Swin transformer-based object detection model using explainable meta-learning mining. *Applied Sciences*, 13(5):3213, 2023. 2

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1

[5] Yuda Bi, Anees Abrol, Zening Fu, and Vince Calhoun. A multimodal vision transformer for interpretable fusion of functional and structural neuroimaging data. *bioRxiv*, pages 2023–07, 2023. 2

[6] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 397–406, 2021. 2

[7] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021. 1, 2

[8] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Explaining transformer-based image captioning models: An empirical analysis. *AI Communications*, 35(2):111–129, 2022. 2

[9] Bowen Dong, Pan Zhou, Shuicheng Yan, and Wangmeng Zuo. Towards class interpretable vision transformer with multi-class-tokens. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 609–622. Springer, 2022. 2

[10] Sofiane Elguendouze, Adel Hafiane, Marcilio CP de Souto, and Anaïs Halftermeyer. Explainability in image captioning based on the latent space. *Neurocomputing*, 546:126319, 2023. 2

[11] Jingzhen He, Junxia Wang, Zeyu Han, Jun Ma, Chongjing Wang, and Meng Qi. An interpretable transformer network for the retinal disease classification using optical coherence tomography. *Scientific Reports*, 13(1):3637, 2023. 2

[12] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. 1

[13] Piotr Komorowski, Hubert Baniecki, and Przemyslaw Biecek. Towards evaluating explanations of vision transformers for medical imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3726–3732, 2023. 2

[14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[15] Zhuang Liu, Yunpu Ma, Matthias Schubert, Yuanxin Ouyang, Wenge Rong, and Zhang Xiong. Multimodal contrastive transformer for explainable recommendation. *IEEE Transactions on Computational Social Systems*, 2023. 2

[16] Yiwei Lyu, Paul Pu Liang, Zihao Deng, Ruslan Salakhutdinov, and Louis-Philippe Morency. Dime: Fine-grained interpretations of multimodal models via disentangled local explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 455–467, 2022. 2

[17] Jie Ma, Yalong Bai, Bineng Zhong, Wei Zhang, Ting Yao, and Tao Mei. Visualizing and understanding patch interactions in vision transformer. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–10, 2023. 2

[18] Rupayan Mallick, Jenny Benois-Pineau, and Akka Zemmari. Ifi: Interpreting for improving: A multimodal transformer with an interpretability technique for recognition of risk events. In *International Conference on Multimedia Modeling*, pages 117–131. Springer, 2024. 2

[19] Evelyn Mannix and Howard Bondell. Scalable and robust transformer decoders for interpretable image classification with foundation models. *arXiv preprint arXiv:2403.04125*, 2024. 2

[20] Arnab Kumar Mondal, Arnab Bhattacharjee, Parag Singla, and A. P. Prathosh. xvitcos: Explainable vision transformer based covid-19 screening using radiography. *IEEE Journal of Translational Engineering in Health and Medicine*, 10:1–10, 2022. 2

[21] Angelos Nalmpantis, Apostolos Panagiotopoulos, John Gkountouras, Konstantinos Papakostas, and Wilker Aziz. Vision diffmask: Faithful interpretation of vision transformers with differentiable patch masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3756–3763, 2023. 2

[22] Usman Naseem, Matloob Khushi, and Jinman Kim. Vision-language transformer for interpretable pathology visual question answering. *IEEE Journal of Biomedical and Health Informatics*, 27(4):1681–1690, 2022. 2

[23] Shami Nisimov, Raanan Y Rohekar, Yaniv Gurwicz, Guy Koren, and Gal Novik. Clear: Causal explanations from attention in neural recommenders. *arXiv preprint arXiv:2210.10621*, 2022. 3

[24] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 1

[25] OpenAi. Gpt-4v(ision) system card. 2023. 1

[26] Bowen Pan, Rameswar Panda, Yifan Jiang, Zhangyang Wang, Rogerio Feris, and Aude Oliva. IA-RED^2 : Interpretability-aware redundancy reduction for vision transformers. *Advances in Neural Information Processing Systems*, 34:24898–24911, 2021. 1

[27] Clément Playout, Renaud Duval, Marie Carole Boucher, and Farida Cheriet. Focused attention in transformers for interpretable classification of retinal images. *Medical Image Analysis*, 82:102608, 2022. 2

[28] Yao Qiang, Chengyin Li, Prashant Khanduri, and Dongxiao Zhu. Interpretability-aware vision transformer. *arXiv preprint arXiv:2309.08035*, 2023. 2

[29] Krithik Ramesh and Yun Sing Koh. Investigation of explainability techniques for multimodal transformers. In *Australasian Conference on Data Mining*, pages 90–98. Springer, 2022. 2

[30] Thomas Richardson and Peter Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030, 2002. 4

[31] Mattia Rigotti, Christoph Miksovic, Ioana Giurgiu, Thomas Gschwind, and Paolo Scotton. Attention-based interpretability with concept transformers. In *International Conference on Learning Representations*, 2022. 1

[32] Raanan Y Rohekar, Shami Nisimov, Yaniv Gurwicz, and Gal Novik. Iterative causal discovery in the possible presence of latent confounders and selection bias. *Advances in Neural Information Processing Systems*, 34:2454–2465, 2021. 3

[33] Raanan Y Rohekar, Yaniv Gurwicz, and Shami Nisimov. Causal interpretation of self-attention in pre-trained transformers. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 4

[34] Debaditya Shome, Tejaswini Kar, Sachi Nandan Mohanty, Prayag Tiwari, Khan Muhammad, Abdullah AlTameem, Yazhou Zhang, and Abdul Khader Jilani Saudagar. Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare. *International Journal of Environmental Research and Public Health*, 18(21):11086, 2021. 2

[35] Jiamei Sun, Sebastian Lapuschkin, Wojciech Samek, and Alexander Binder. Explain and improve: Lrp-inference fine-tuning for image captioning models. *Information Fusion*, 77: 233–246, 2022. 2

[36] Vinitra Swamy, Malika Satayeva, Jibril Frej, Thierry Bossy, Thijs Vogels, Martin Jaggi, Tanja Käser, and Mary-Anne Hartley. Multimodn—multimodal, multi-task, interpretable modular networks. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[37] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

[38] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*, 2024. 4

[39] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[40] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[42] Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023. 1

[43] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019. 1

[44] Tianfu Wu and Xi Song. Towards interpretable object detection by unfolding latent structures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6033–6043, 2019. 2

[45] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024. 1

[46] Mengqi Xue, Qihan Huang, Haofei Zhang, Lechao Cheng, Jie Song, Minghui Wu, and Mingli Song. Protopformer: Concentrating on prototypical parts in vision transformers for interpretable image recognition. *arXiv preprint arXiv:2208.10431*, 2022. 2

[47] Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008. 4

[48] Quanshi Zhang, Ruiming Cao, Feng Shi, Ying Nian Wu, and Song-Chun Zhu. Interpreting cnn knowledge via an explanatory graph. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 1

[49] Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. Interpreting cnns via decision trees. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6261–6270, 2019. 1

[50] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024. 1