IDENTIFYING CHALLENGES FOR GENERALIZING TO THE PEARL CAUSAL HIERARCHY ON IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Towards the ultimate goal of AI that features agents capable of generalizing to unseen domains, many researchers have recently voiced their support towards Pearl's counterfactual theory of causation as a key milestone. As in any other growing subfield, patience seems to be a virtue since significant progress on integrating notions from both fields takes time, yet, major challenges such as the lack of ground truth benchmarks or a unified perspective on classical problems such as computer vision seem to hinder the momentum of the research movement. This work takes a first, informal look at the Pearl Causal Hierarchy (PCH) for image data. We moreover discuss several challenges that naturally arise when applying key concepts from causality to the study of image data.

1 INTRODUCTION AND RELATED WORK

The Pearlian counterfactual theory of causation (Pearl, 2009) has increasingly found support in the AI/ML community (Schölkopf, 2022; Peters et al., 2017; Geffner et al., 2022). An increasing presence of publications at major conferences/journals concerned with the integration of causality with AI/ML (including (Janzing & Schölkopf, 2018; Lee & Bareinboim, 2019; Zečević et al., 2021) to mention a select few), but also the establishment of new conferences such as CLeaR, suggests a growing subfield that sets a consensus on *causal* AI/ML as answer to question "what do we need for successful domain generalization?". Still, as the difficulty of the integration with otherwise prominent success stories of deep learning such as computer vision becomes apparent, countering opinions start speaking out against causal AI/ML (Bishop, 2021). Nonetheless, we take the arguably agreed upon perspective *pro* causal AI/ML and we specifically try addressing challenges that arise when viewing computer vision from a causal viewpoint.

Naturally, we are not the first to discuss causality in terms of computer vision. Several works at popular venues such as CVPR (including (Sauer & Geiger, 2021; Lv et al., 2022; Liu et al., 2022) to mention a select few) have taken on the challenge. Most notably, Sanchez & Tsaftaris (2022) looked at both diffusion models and causality, like this work, however, with the goal of counterfactual image generation opposed to a consistent interpretation of the PCH in terms of images.

In Sec.2 we will present a first, informal interpretation of the PCH on image data, that is, for each of the levels of the hierarchy, what would corresponding images look like. In Sec.3 we discuss four different challenges to the perspective discussed prior that might come as surprising and that pose difficulties to finding models that can successfully generate a correct hierarchy on image data. Lastly, in Sec.4 we provide some last reflective thoughts on the presented introspection of the community's progress towards generalization using causal computer vision. We also provide an appendix with additional considerations for potential benchmark tasks of interest for evaluating future models.

2 A FIRST INTERPRETATION OF THE CAUSAL HIERARCHY ON IMAGES

The center model of study in causality is the *structural causal model* (SCM) which is defined as a 4-tuple $\mathcal{M} := \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$ where the so-called structural equations (which are deterministic functions) $v_i = f_i(\mathrm{pa}_i, u_i) \in \mathcal{F}$ assign values (denoted by lowercase letters) to the respective endogenous/system variables $V_i \in \mathbf{V}$ based on the values of their parents $\mathrm{Pa}_i \subseteq \mathbf{V} \setminus V_i$ and the values of some exogenous variables $\mathbf{U}_i \subseteq \mathbf{U}$ (sometimes also referred to as *unmodelled* or

nature terms), and $P(\mathbf{U})$ denotes the probability function defined over $\mathbf{U}$. The SCM formalism comes with several interesting properties. They induce a causal graph G and they induce what is known as the *Pearl Causal Hierarchy* (PCH). Pearl & Mackenzie (2018) provide an intuitive account to the topic. The hierarchy (sometimes also referred to as ladder) consists of three levels. The first, $\mathcal{L}_1$, is about observational/associational distributions over $\mathbf{V}$ with typical questions being "What is?", for example "What does the symptoms tell us about the disease?". While on $\mathcal{L}_1$ we are only concerned with a single observational distribution, on the second and the first causal level, $\mathcal{L}_2$, we have infinitely many interventional/hypothetical distributions with typical questions being "What if?", for example "What if I take an aspirin, will my headache be cured?". Finally and again infinitely many, counterfactual/retrospective distributions can be found on the third level, $\mathcal{L}_3$, with typical questions being "Why?", for example "Was it the aspirin that cured my headache?". A key result and sort of "sanity check" for research in causality was the establishment of the *Causal Hierarchy Theorem* (Bareinboim et al., 2020) which suggests (a) that any SCM will imply the PCH as just discussed with its $\mathcal{L}_1$ *associational*, $\mathcal{L}_2$ *interventional* and $\mathcal{L}_3$ *counterfactual* levels, and more fundamentally (b) that causal quantities ($\mathcal{L}_i, i \in \{2, 3\}$) are in fact richer in information than statistical quantities ($\mathcal{L}_1$), and that there exists a necessity of causal information (e.g. structural knowledge, essentially "outside" model knowledge) for inference based on lower rungs e.g. $\mathcal{L}_1 \not\Rightarrow \mathcal{L}_2$ and therefore to reason about $\mathcal{L}_2$ or to *identify* such causal quantities we need more than only observational data from $\mathcal{L}_1$. To conclude, consider the formal definition of valuations for the highest layer ($\mathcal{L}_3$) since it subsumes the other two layers as previously pointed out:

$$p(\mathbf{a_b}, \ldots, \mathbf{c_d}) = \textstyle\sum_{\mathcal{U}} p(\mathbf{u}) \quad \text{where} \quad \mathcal{U} = \{\mathbf{u} \mid \mathbf{A_b}(\mathbf{u}) = \mathbf{a}, \ldots, \mathbf{C_d}(\mathbf{u}) = \mathbf{c}\},$$

for instantiations $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ of the node sets $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D} \subseteq \mathbf{V}$ and they represent different "worlds". The counterfactual $\mathbf{A_b}(\mathbf{u})$ refers to the value $\mathbf{A}$ attains when $\mathbf{B}$ was deliberately *set* to $\mathbf{b}$ in situation $\mathbf{u}$. E.g. for $\mathcal{L}_1$ we might only consider $\mathbf{A} = \mathbf{A}_\emptyset$, whereas for $\mathcal{L}_2$ a single alternate world $\mathbf{A_b}$. Next, we define for the first time informally the $\mathcal{L}_i$ for images. We start with the standard case $\mathcal{L}_1$, then $\mathcal{L}_3$ as it is easier to develop from the previous and then conclude with $\mathcal{L}_2$. We summarize the insights collected in this section in a comprehensive schematic in Fig.1 that highlights both the hierarchical nature and the requirements for commuting between the different levels.

$\mathcal{L}_1$ on Images. In computer vision, we work with image data which is naturally represented as an ordered collection of pixel values commonly represented as matrices with multiple channels (e.g. RGB). Now, a common way of linking images with causality's SCM has been to have the SCM act as the image generating process. That is, the SCM will not represent pixels but rather "high level concepts" that nicely abstract the content and style of the images at hand. E.g. $\mathbf{V}$ could be specifying concepts like 'dog,' which could be an indicator function suggesting that a 'dog' should be placed within the image generated through that concrete specification. However, it is important to note that the SCM, as usual, is the sought after object of interest i.e., we do not know the model yet. In that sense, Pearlian causality comes in handy as a tool for abstraction and simply a language for *formalizing* the modelling assumptions. The first question we will answer now is, how do observations on image data, that is, instances of $\mathcal{L}_1$ for images look like? The bottom image in Fig.1 illustrates an example. We can see a dog facing the image observer with its mouth open, tongue sticking out, ears upright, and sitting on a dry patch of lawn. The dog's fur is orange and white colored, the race of the dog is classified as corgi. This short description are all 'objective' observations using human concepts about the appearance of the image that might or might not be captured formally in a corresponding SCM. Generally, we can state that $\mathcal{L}_1$ on images will simply correspond to any image collection that we consider to be our base set of images, however, with the important restriction that 'related' images are excluded. As we will see shortly, a random set of images might very well contain such related images, which will naturally be a mix of $\mathcal{L}_{1/2}$ data.

$\mathcal{L}_3$ on Images. What happens if you take the image of the corgi from before and *change* aspects to it? Contrary to expectations, we end up on the final level of the PCH through an intervention on the *same image*. This might be counterintuitive and surprising since data augmentations are common practice in computer vision and usually associated with pure interventions ($\mathcal{L}_2$). Ilse et al. (2021) discussed one such interpretation of augmentations in a causal setting. Using modern techniques from deep learning like stable diffusion (Rombach et al., 2022) and/or a combination of inpainting and image editing we can generate a counterfactual image (countering the fact that we initially observed only the corgi) where there is suddenly a red apple placed right in front of the corgi on the lawn. In terms of an SCM that captures high level concepts on content, we can say that the

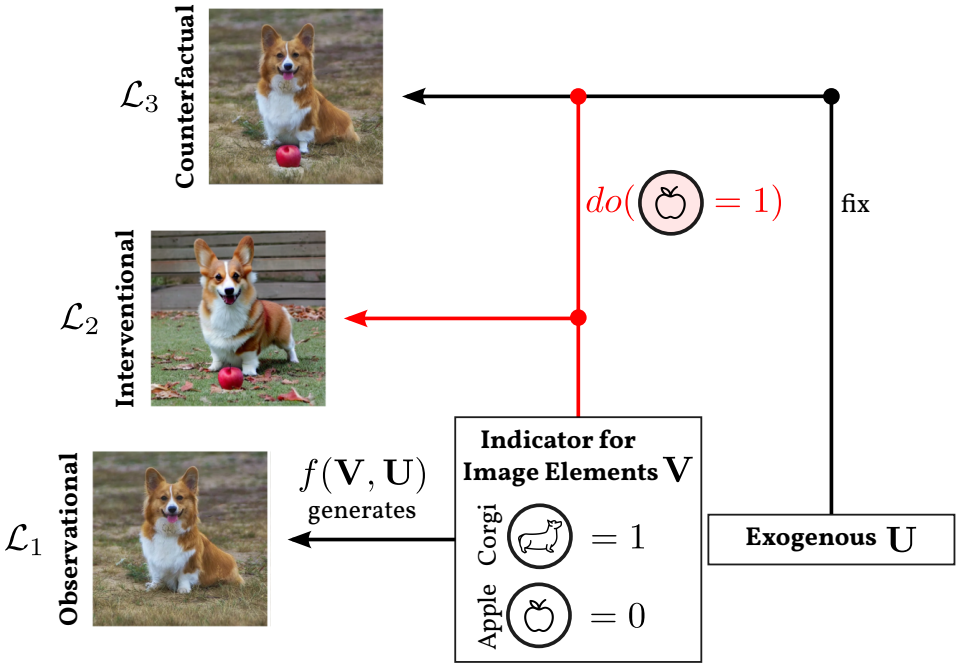

Figure 1: **Pearl Causal Hierarchy on Image Data.** In the first level, $\mathcal{L}_1$, we are given some base image $\mathbf{X}_{\mathcal{L}_1}$ like the one shown that pictures a corgi facing the viewpoint. We can assume the SCM that generated this base image to capture high level binary indicators such as V_1 :="corgi in image" or V_2 :="apple in image" and for the base image, therefore, we observed $\mathbf{V} = (1, 0)$. Naturally, unmodelled 'details' $\mathbf{U}$ are concerned with specifications on what the corgi should look like, its pose, but also the background amongst others. Therefore, we have $\mathbf{X}_{\mathcal{L}_1} = f(\mathbf{V}, \mathbf{U})$. To jump to the second level, $\mathcal{L}_2$, we can simply change $\mathbf{V}$ by intervening with $do(V_2 = 1)$ for instance, which leads to the appearance of a red apple in the intervened image $X_{\mathcal{L}_2}$. Note how the background and the corgi itself change through the intervention, this is because there is no restrection on the details $\mathbf{U}$ being placed by the intervention. That is, $\mathbf{U}$ is different for $X_{\mathcal{L}_1}$ and $X_{\mathcal{L}_2}$. Finally, on the third level, $\mathcal{L}_3$, we can re-introduce that restriction by fixing $\mathbf{U}$ and doing the intervention to return an image which is like $X_{\mathcal{L}_1}$ up to the intervention of placing the apple. (Best viewed in color.)

exogenous terms $\mathbf{U}$ used for generating the actual images from the content indicators $\mathbf{V}$ are being kept fixed while the $\mathbf{V}$ has been intervened upon such that the "apple indicator" is flipped on.

$\mathcal{L}_2$ on Images. To now obtain pure interventions, we need to relax the constraint that the new image would have to fix the $\mathbf{U}$ terms, that is, all the noise apart from the content indicators. Put simply, we don't have to observe the original, base image anymore. However, we still need to observe content for our given SCM in agreement with the base image, for example in this case a corgi should still be found in the intervened image. An example intervention is shown in the image on the right, which again presents a corgi and a red apple, but a different corgi on a different background thus not being constrained by the 'factual' original image on all the unmodelled 'details' of the image. We note two observations, (1) that the intervened image is difficult to synthesize through manual labour, that is, diffusion models open a new avenue of opportunity for creating true interventional image distributions, and (2) that the counterfactual from $\mathcal{L}_3$ can really be seen as a special case of intervened images, the one that agrees with the base image on $\mathbf{U}$.

3 Challenges

After our initial discussion on how one can interpret the PCH in terms of image data, exemplifying how images corresponding to each of the levels might look, we now move on to the finer details that come about naturally when being concerned with images, which complicate the overall problem of causal inference on images.

Challenge 1: Ambiguity of Interventions. We have seen how interventions can be naturally formulated on the *content* abstraction of images, that is, if there are high level concept indicators such as "is there a dog in the image?", then answering 'yes' or 'no' has predictable consequences in that a dog will or will not appear in the resulting image. However, what happens if we intervene on not what is being placed in the image but rather what is already within the image like, say, a bird sitting on a tree branch? If we do something like "spread the wings of the bird", then we surely underspecify what we mean by "spreading wings."

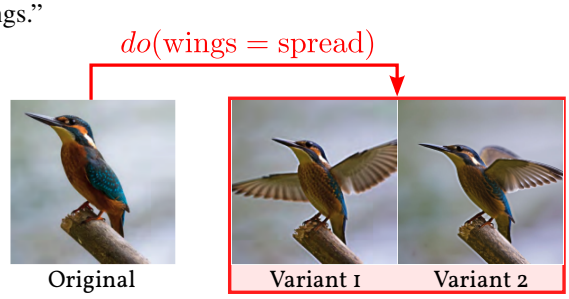

Figure 2. The figure on the right illustrates this idea that an intervention, especially in textual form as just given in the example, leaves open several aspects that somehow need to be decided for in the final image. In this example, the bird might spread the wings at an angle or to a certain range of motion as illustrated in variants 1 and 2, respectively. This observation becomes very apparent with cutting-edge methods in diffusion modeling currently under peer-review like Imagic (Kawar et al., 2022) or Unitune (Valevski et al., 2022).

Challenge 2: Different Types of Interventions. In line with the discussion in the previous subsection, we could already see how "not all interventions are created equal" in that interventions can be of *qualitatively* different nature.

Figure 3. To be more precise, in the corgi example from Sec.2 we simply placed an apple in front of the corgi, whereas in the bird example we let the bird in the image spread its wings. Consider the figure on the right that places the examples on top of each other. The intervention in the top row places a new object within the image and naturally it also changes aspects of the previous image, for instance we cannot see the parts of the corgi or lawn that is being occluded by the newly placed apple. The intervention in the bottom row changes the state of the intervened variable, here the bird, in a way that is less local than previously in the object placement intervention. While it is difficult to capture the essence of what makes these interventions fundamentally different even in informal terms, we believe the main distinction to be that the object placement poses a pure intervention *on the image*, whereas the wings spreading poses an intervention *on the image elements*. This distinction we believe to be important because the latter contains a notion of *implicit causation*, which can be seen as an indicator of the 'plausibility' behind the image, for instance spreading the wings involves changing the bird's posture, the position of its limbs etc. To make this point more clear, imagine we placed a bone in the mouth of the corgi in the first example, this would constitute at first sight an independent object placement intervention, however, ideally, we expect the dog to hold it in its mouth by biting on it, which implies a causality similar to the second example with the bird. In any case, many of these aspects are up for debate, more importantly we wish to emphasize that the term 'intervention' could already be ill-posed w.r.t. image data.

Challenge 3: Inpainting versus Fine-Tuning. This challenge is concerned with two existing methods with which interventional and counterfactual images can be generated right now: inpainting (IP) and fine-tuning of existing diffusion models (FT). In IP we generally have two options we can either mask out certain image segments like for instance the mouth of the corgi from the first example and then command a diffusion model to fill in the newly created gap by "placing a bone in the corgi's mouth", or we can directly use photo editing software to modify the original image. In FT on the other hand, we might deploy an approach like Imagic (Kawar et al., 2022) or Unitune (Valevski et al., 2022) where the authors fine-tune existing diffusion models to regularize the embedding space in such way that for any given intervention the synthesized images adhere more strongly to what is being asked for by the intervention. For example in a method such as SDEdit (Meng et al., 2021) or Text2Live (Bar-Tal et al., 2022) the synthesized images might ignore the intervention altogether, or even worse corrupt the base image. Both IP and FT are techniques readily available for reproduction

and could thus be deployed to generate a *truly* causal data set consisting of images from all three levels of the PCH (even if models like Imagic and Unitune are unfortunately being kept disclosed under intellectual property rights). However, such data set generation is practically still infeasible since both IP and FT are *costly* methods for generating causal data, especially at the scale of modern techniques of deep learning. For IP, either manual labour or the restriction to certain types of interventions (like object placement) would be necessary to synthesize interventional and counterfactual images, while for FT, fine-tuning for each generation would be necessary, thereby rendering both approaches practically irrelevant for synthesizing causal image data sets.

Challenge 4: Inaccuracy of Counterfactuals. In this last subsection on intricate challenges we discuss how, not only interventions, but also counterfactuals introduce issues that complicate causal inference on images. A counterfactual, as defined by Pearl, subsumes a three step procedure that involves updating your beliefs on the exogenous $\mathbf{U}$ but also performing interventions. We have seen ambiguities caused by interventions but similar ones also come with the exogenous 'details' that constitute part of any SCM.

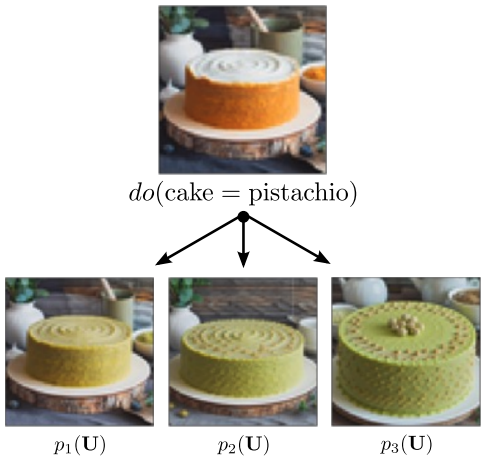

Figure 4. Consider the example given in the figure on the right, where the base image of a cake is being intervened upon such that the new image contains an image of a pistachio flavored cake. Ideally, the counterfactual should be the same as the base image up to the intervention. Yet, as we realize from our previous discussions, what constitutes 'same' is not clear, that is, where do we *exactly* find all the aspects of the exogenous terms within the image? The example highlights three different distributions over $\mathbf{U}$, representing different beliefs (or conceptions), of the exogenous terms and what 'same' *ought* to mean. We can clearly see that there is no ideal answer i.e., out of all three distribution $p_1(\mathbf{U})$ keeps the surrounding most constant (e.g. by looking at the wooden serving plate and the similar viewpoint) but is arguably also the "least pistachio" like out of the three. Same arguments can be made for the other two, e.g. $p_3(\mathbf{U})$ even contains pieces of pistachio on top of the cake but the background changed significantly introducing new dishware and also a very different viewing angle. Looking at $p_2(\mathbf{U})$ for instance then seems to compromise between the two previous but is significantly wider than the cake from the base image. The highlighted issue of inaccuracy in counterfactuals becomes especially prevalent in methods such as Imagic (Kawar et al., 2022) that make use of thresholding to simply settle for a decided notion of proximity, which further is likely to require human supervision to ensure the image's integrity and purpose. To move forward, we believe that any of three beliefs about what constitutes $\mathbf{U}$ would have been sufficient for most relevant tasks of interest, nonetheless, we intend on raising awareness to this intricacy which might otherwise would have fallen under the radar, since the ambiguity otherwise naturally found in images also naturally translates to a proximity issue for counterfactual images.

4 Concluding Remarks

The belief in the utility of causality for AI/ML permeates the ongoing narrative, uniting the community in the hope that causal models hold a significant part of the future of intelligent systems. However, there is still a bitter taste left for many, since they do agree on the fact that being able to answer causal queries of the interventional and counterfactual nature is desirable, but worry about the promise never being delivered. This feeling is only being further corroborated by a lack of existing data sets, especially in domains that have historically been the success stories of AI/ML like computer vision. But also through a lack of clarity in how the intersection of causality and computer vision is to be understood. In this work we highlighted how we can interpret and project the Pearl Causal Hierarchy on image data and identified challenges that come with it, therefore, providing a clear perspective on what we might need for successful domain generalization in computer vision for image sets that share causal relations. We hope that this work can inspire researchers by painting a vivid picture of key ideas and emphasizing important avenues for future work.

REFERENCES

Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. *arXiv preprint arXiv:2204.02491*, 2022.

Elias Bareinboim. Tutorial on "towards causal reinforcement learning", 2020. URL `https://www.youtube.com/watch?v=QRTgLWfFBMM`.

Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. 1on pearl's hierarchy and. Technical report, Technical Report, 2020.

J Mark Bishop. Artificial intelligence is stupid and causal reasoning will not fix it. *Frontiers in Psychology*, 11:2603, 2021.

Hector Geffner, Rina Dechter, and Joseph Y Halpern. Probabilistic and causal inference: The works of judea pearl, 2022.

Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. *Advances in Neural Information Processing Systems*, 32, 2019.

Maximilian Ilse, Jakub M Tomczak, and Patrick Forré. Selecting data augmentation for simulating interventions. In *International Conference on Machine Learning*, pp. 4555–4562. PMLR, 2021.

Dominik Janzing and Bernhard Schölkopf. Detecting non-causal artifacts in multivariate linear regression models. In *International Conference on Machine Learning*, pp. 2245–2253. PMLR, 2018.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022.

Sanghack Lee and Elias Bareinboim. Structural causal bandits with non-manipulable variables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4164–4172, 2019.

Weiyang Liu, Zhen Liu, Liam Paull, Adrian Weller, and Bernhard Schölkopf. Structural causal 3d reconstruction. In *European Conference on Computer Vision*, pp. 140–159. Springer, 2022.

Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8046–8056, 2022.

Aengus Lynch, Jean Kaddour, and Ricardo Silva. Evaluating the impact of geometric and statistical skews on out-of-distribution generalization performance. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10971–10980, 2020.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Pedro Sanchez and Sotirios A Tsaftaris. Diffusion causal models for counterfactual estimation. In *Conference on Causal Learning and Reasoning*, pp. 647–668. PMLR, 2022.

Axel Sauer and Andreas Geiger. Counterfactual generative networks. *arXiv preprint arXiv:2101.06046*, 2021.

Bernhard Schölkopf. Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 765–804. 2022.

Sahil Singla and Soheil Feizi. Causal imagenet: How to discover spurious features in deep learning? *arXiv preprint arXiv:2110.04301*, 2021.

Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. *arXiv preprint arXiv:2210.09477*, 2022.

Matej Zečević, Devendra Dhami, Athresh Karanam, Sriraam Natarajan, and Kristian Kersting. Interventional sum-product networks: Causal inference with tractable probabilistic models. *Advances in Neural Information Processing Systems*, 34, 2021.

A APPENDIX
"TASKS OUR FUTURE MODELS NEED TO SOLVE"

Selected, additional discussions as supplementary and optional read to the main paper.

In the previous section we discussed several challenges to causal inference with images, but now we turn our attention to actual challenges in the sense of *tasks that we wish our models to be capable of solving*. We present two different tasks that arise naturally as a consequence of our interpretation of the PCH for image data.

Previous Attempts at Causal Image Data Sets. There have been several strides in coming up with a causal or 'kind-of' causal data set. For instance, Gondal et al. (2019) looked at disentanglement and factors of variation and created both synthetic and real-world images of objects in different poses, colors, etc. Singla & Feizi (2021) used manual labour on mechanical turk to study spurious features in images, while Lynch et al. (2022) recently employed diffusion models and base assumptions on the underlying SCM to synthesize data sets that balance out each of the existing high level concepts (e.g. if there was never a cow on the beach in the original data set, then now there was). All of these aforementioned approaches mitigate some of the core issues with causal data sets, especially for images. However, none of them have been able to fully subsume the PCH, and unfortunately, due to the difficulties with existing techniques as highlighted in Sec.3, a truly causal image data set seems to still be out of reach as not even brute force techniques could force enough data samples for modern ML techniques lingering for ever-increasing data amounts. Nonetheless, we believe that such a data set would constitute only part of the overall big picture. What do we do when given such a 'truly' causal image data set? Next we try providing an answer to this question by highlighting tasks for our future models.

A.1 TASK I: IMAGINING A COUNTERFACTUAL IMAGE

The task is being illustrated schematically in Fig.5 with the bird example from the main paper discussions. Counterfactuals pose, in Pearl's term, the center piece to his theory of causation and are correspondingly placed on the highest level (or rung) of the PCH. Opposed to pure interventions which talk about hypothetical situations, counterfactuals are all about *retrospection*. That is, through knowing about the exogenous terms $\mathbf{U}$, which in the standard case imply a certain $\mathbf{V} = \mathbf{v}$ since fixing $\mathbf{U} = \mathbf{u}$ results in deterministic functions $\mathbf{v} = f(\mathbf{pa}, \mathbf{u})$, we move "back in time" and consider an alternate *world* configuration. This observation also gives rise to Pearl's opinion that counterfactuals corresponding to 'understanding' and the "highest mode of human cognition" (Pearl & Mackenzie (2018) further illustrate $\mathcal{L}_3$ with a cartoon sketch of Einstein reflecting on complex inventions such as rockets or laptops). As of now, diffusion models are capable of creating counterfactuals (as clearly the given examples in this paper illustrate, since they were indeed generated with the help of diffusion models), however, they require fine-tuning or inpainting, which effectively changes the original model (which amounts to defeating the purpose of the challenge). Therefore, since counterfactual reasoning is desirable, and current models are not able to generate counterfactuals in a *targeted manner*, we propose in Task I to "imagine counterfactuals." That is, given an input tuple containing an image and an intervention, synthesize the corresponding counterfactual image. In order to succeed in this task the queried system needs to recognize the point of intervention (here: bird, and that particular bird in the cases that there are multiple birds) and also 'know' what the consequences of the given intervention are (here: limbs moving, posture changing etc. due to the wing spread).

A.2 TASK II: IDENTIFYING THE INTERVENTIONS WITHIN IMAGES

The intricate relation between interventions and counterfactuals becomes more apparent in the second task that we propose, which is illustrated in Fig.6, again with the bird example. Forming a counterfactual involves performing an intervention in a fixed-$\mathbf{U}$ world. The task is specified as follows: given an input tuple containing a base image and a corresponding image in which an intervention occurred, extract the corresponding intervention. The intervention is implicitly hidden within the image and the queried system needs to be able to identify the change between the images and explicate the *reason* for the change (here: the bird spreading its wing). While in Task I the queried system is concerned with the same hypothesis space (that of images) for providing the

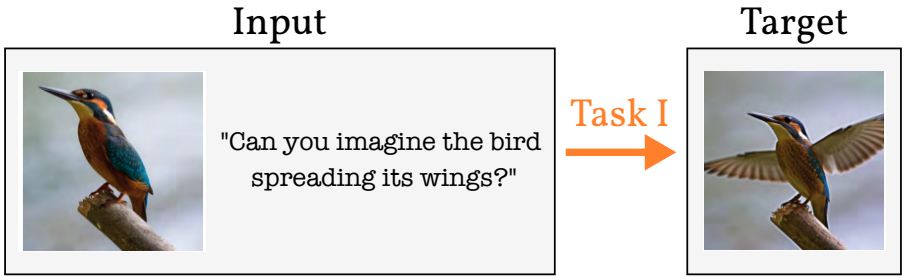

Figure 5: **Task I, "Imagining Counterfactuals."** Given an input tuple containing an image and an intervention, *synthesize* the corresponding counterfactual image. (Best viewed in color.)

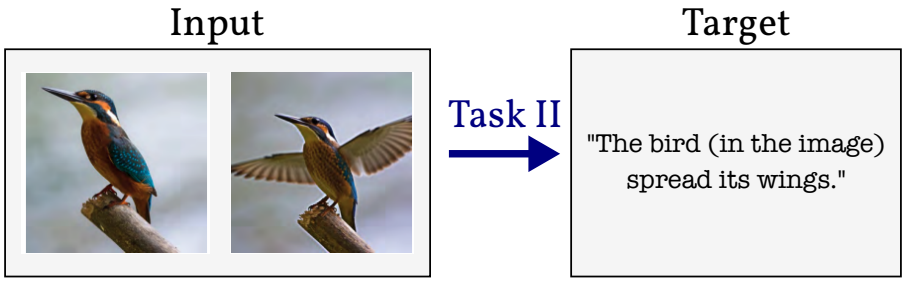

Figure 6: **Task II, "Identifying Interventions."** Given an input tuple containing a base image and a corresponding image in which an intervention occurred, *extract* the corresponding intervention. (Best viewed in color.)

answer, in Task II the multi-modality switches the answering space to be in natural language. Again, like with Task I, existing state-of-the-art methods for image captioning like X-LAN (Pan et al., 2020) are capable of capturing precisely what is "going on" in a given image, and thus a naïve solution to Task II would be to simply look at some sensible notion of difference between the two independently captioned images. However, we'd again be defeating the purpose since we want to have an *inherently single* model being capable of solving this task by processing the input pair through some internal representation in such way that the intervention becomes apparent.

Solving Tasks I and II. If we had a model that can solve Tasks I and II, then we'd have a model that can (a) understand interventions and (b) use them to reason counterfactually–and all of that within images. Nonetheless, a very important aspect to this, that fell short in the previous discussions, is the aspect of time or dynamics unrolling. We humans know how it looks like when a bird is spreading its wings, and so we are able to solve both Tasks I and II gracefully although being confronted with *still* images (with the asteriks on Task I because a human, without aid, will not provide the pixel values for the counterfactual image). We learn to explore the world interactively in a Markov Decision Process type situation (see for instance the Bareinboim (2020) ICML Tutorial on Causal Reinforcement Learning). Therefore, we need to be aware of the assumptions and restrictions we place on the learning problem for our models when facing Tasks I and II.