VALIDCF: EVALUATING SYNTHETIC COUNTERFACTUAL EVIDENCE IN MEDICAL VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthetic counterfactuals are increasingly used to evaluate medical vision-language models (VLMs). They test whether a model responds to clinically meaningful visual changes. However, the editing process can also introduce unintended changes. This creates a fundamental question: Did the model respond to the intended clinical change or to the editing process itself? We introduce ValidCF, a framework for evaluating synthetic counterfactual evidence before using it to interpret medical VLM behavior. ValidCF pairs each clinical counterfactual with a matched sham intervention. The sham follows the corresponding editing pipeline while preserving the target clinical state. This provides a control for measuring editing-associated response instability. ValidCF also examines intervention-validity evidence and consistency across independent counterfactual generators. Uncertain interventions can then be accepted, deferred, or rejected rather than treated as equally reliable evidence. We evaluate ValidCF on 500 patient-distinct chest radiographs using two independent counterfactual generators and a medical VLM. We find a striking disagreement between conventional counterfactual evaluation and sham-controlled analysis. The generator with higher clinical transition accuracy also produces substantially greater response changes under sham interventions. Once this instability is considered, the apparent comparison between the generators reverses. Our findings show that an expected response to a counterfactual edit is not, by itself, sufficient evidence of clinical sensitivity. The intervention used to produce that evidence must also be evaluated. ValidCF provides a controlled approach for distinguishing intended clinical responsiveness from sensitivity to the editing process.

1 INTRODUCTION

Medical vision-language models (VLMs) are increasingly evaluated beyond answer accuracy. A correct answer does not necessarily show that a model relied on the visual evidence that should support it. Medical VLMs can exploit linguistic or visual shortcuts, motivating evaluations of whether predictions are grounded in clinically relevant image evidence Nguyen et al. (2025); Xiong et al. (2026).

Counterfactual interventions provide a natural way to probe this problem. They test whether model behavior changes when clinically meaningful evidence changes. Recent methods can add, remove, or modify findings in chest radiographs while attempting to preserve unrelated content Pérez-García et al. (2024); Min et al. (2025); Cooke et al. (2026). Counterfactuals have also been used to expose confounding and test visual attribution Fathi et al. (2024); Zhang et al. (2025); Xiong et al. (2026).

This approach, however, places considerable trust in the intervention itself. A synthetic counterfactual is not the observed alternative state of the same patient. An editor may modify the target finding while also changing non-target image properties. Existing methods therefore constrain unintended modifications, preserve non-target information, or evaluate target effectiveness alongside unintended effects Min et al. (2025); Cooke et al. (2026); Stanley et al. (2025). More fundamentally, counterfactual image editing requires causal assumptions that cannot generally be identified from observational image–label pairs alone Pan & Bareinboim (2024).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

These advances leave a related question about the downstream model response:

> *If a medical VLM changes its answer after a synthetic counterfactual edit, did it respond to the intended clinical change or to the editing process itself?*

We introduce **ValidCF** to examine this ambiguity. ValidCF pairs each clinical counterfactual with a matched sham intervention that preserves the target clinical state. The sham acts as a negative control for editing-associated response instability Lipsitch et al. (2010). A response change under the sham condition shows that the editing pipeline can alter VLM behavior without the intended clinical transition. We further compare independent counterfactual generators to examine dependence on the intervention mechanism. These controls do not identify the true causal counterfactual. Instead, they test the stability of the evidence used to infer clinical sensitivity.

We evaluate ValidCF on 500 patient-distinct chest radiographs using two counterfactual generators and a medical VLM. InstructX2X achieves higher clinical transition accuracy than RadEdit (73.4% vs. 67.0%), but also substantially higher sham response instability (44.8% vs. 11.6%). Their matched response contrasts are consequently 28.6% and 55.4%, respectively. Thus, the generator favored by target-transition accuracy does not provide the stronger separation between intended clinical response and sham instability. This disagreement shows why target-transition accuracy alone can provide an incomplete account of counterfactual evidence sensitivity.

Our contributions are threefold. First, we introduce a matched sham-control protocol for detecting editing-associated response instability in medical VLM counterfactual evaluation. Second, we show that clinical transition accuracy and sham stability can support conflicting conclusions across counterfactual generators. Third, we distinguish *intervention uncertainty* from prediction uncertainty and use a three-way decision mechanism to Accept, Defer, or Reject synthetic counterfactual evidence.

2 Related Work

Visual grounding and counterfactual medical imaging. Medical VLM evaluation increasingly asks whether correct answers are supported by relevant visual evidence. HEAL-MedVQA tests textual and visual shortcuts Nguyen et al. (2025), while MedVIGIL examines model behavior when supporting visual evidence is compromised Jiang et al. (2026). MedFocus uses targeted counterfactual editing to test visual attribution in chest X-ray reasoning Xiong et al. (2026). These studies distinguish answer accuracy from evidence grounding. We examine a complementary question: whether the synthetic intervention used to test grounding is itself sufficiently reliable to support the resulting conclusion.

Medical counterfactual methods attempt to modify a target finding while preserving unrelated information. RadEdit uses diffusion-based editing with explicit edit and preservation masks Pérez-García et al. (2024); InstructX2X performs localized instruction-driven editing Min et al. (2025); and RoentMod modifies specified chest X-ray findings while preserving patient characteristics Cooke et al. (2026). Intervention validity nevertheless remains challenging. DeCoDEx constrains confounding during counterfactual generation Fathi et al. (2024), while Stanley et al. evaluate intended effects together with unintended changes Stanley et al. (2025). More fundamentally, counterfactual image editing is generally not identifiable from observational image–label data alone Pan & Bareinboim (2024). ValidCF focuses on the downstream consequence: when an edited image is used to evaluate a VLM, the intervention becomes part of the measurement process.

Sham controls and intervention mechanisms. Negative controls are used to detect bias when the target effect should be absent Lipsitch et al. (2010). We apply this principle through target-preserving sham interventions. Related null-edit designs also appear in EEG-EditBench Zhang et al. (2026); therefore, our contribution is not null editing itself, but using matched sham responses to evaluate the stability of synthetic counterfactual evidence for medical VLMs.

The intervention mechanism may also influence the measured response. Work on synthetic images shows generator-dependent signatures and sensitivity to generator diversity Zhu et al. (2023); Park & Owens (2025). Related counterfactual methods seek robustness to nuisance variation Chang et al. (2021); Roschewitz et al. (2025); Zhang et al. (2025). Rather than detecting the generator, we

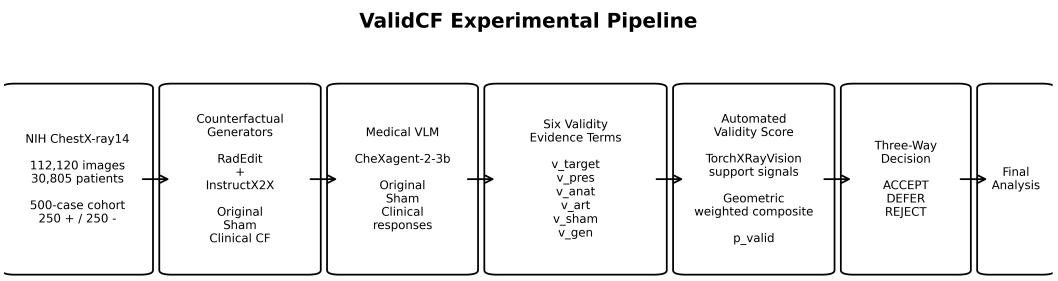

Figure 1: **ValidCF framework.** Clinical and target-preserving sham interventions test expected VLM transitions and editing-associated response instability. Automated intervention evidence is summarized by S_{IV} and mapped to Accept, Defer, or Reject.

ask whether conclusions about VLM evidence sensitivity change across independently developed intervention mechanisms.

Three-way decisions for intervention uncertainty. Three-way decision theory extends binary acceptance and rejection with a boundary region for noncommitment or deferment Yao (2010; 2011). ValidCF applies this principle to the *intervention*, not the medical prediction. A VLM may be confident while the synthetic evidence used to evaluate it remains uncertain. We therefore distinguish

$$\text{prediction uncertainty} \neq \text{intervention uncertainty}.$$

ValidCF connects these previously separate directions through one question: *when should a synthetic counterfactual be trusted as evidence about a medical VLM's clinical sensitivity?* We address this using matched sham controls, cross-generator evidence, and three-way intervention decisions.

3 METHODOLOGY

3.1 VALIDCF FRAMEWORK

Figure 1 summarizes the ValidCF workflow. Starting from an original image, each generator produces a clinical counterfactual and a target-preserving sham intervention. Their responses and automated evidence are then combined to evaluate the intervention before it is used to interpret VLM behavior.

Let $\mathcal{D} = \{(I_i, q_i, a_i)\}_{i=1}^{N}$ denote medical images, clinical questions, and expected answers, and let $f_\theta(I_i, q_i)$ denote the prediction of a medical VLM. For a target intervention z_i, generator G_g produces a clinical counterfactual

$$I_i^{\mathrm{cf},g} = G_g(I_i, z_i^{\mathrm{cf}}), \tag{1}$$

with expected answer a_i^{cf}. Conventional evaluation asks whether $f_\theta(I_i^{\mathrm{cf},g}, q_i) = a_i^{\mathrm{cf}}$. This does not establish that the response is specific to the intended clinical change because the editing process may also modify non-target image properties.

ValidCF therefore pairs each clinical counterfactual with a target-preserving sham,

$$I_i^{\mathrm{sham},g} = G_g(I_i, z_i^{\mathrm{sham}}). \tag{2}$$

The sham provides a negative-control condition in which the intended clinical transition is absent. A response change under this condition reveals editing-associated response instability. It does not recover the true patient counterfactual or remove all generator-specific confounding.

3.2 SHAM-CONTROLLED VLM RESPONSE

We measure three complementary aspects of VLM behavior:

$$\mathrm{SFR}_g = \frac{1}{N}\sum_i \mathbb{I}\Big[f_\theta(I_i^{\mathrm{sham},g}, q_i) \neq f_\theta(I_i, q_i)\Big], \tag{3}$$

$$\mathrm{CTA}_g = \frac{1}{N}\sum_i \mathbb{I}\Big[f_\theta(I_i^{\mathrm{cf},g}, q_i) = a_i^{\mathrm{cf}}\Big], \tag{4}$$

$$\mathrm{MRC}_g = \mathrm{CTA}_g - \mathrm{SFR}_g. \tag{5}$$

Sham Flip Rate (SFR) measures response instability under a target-preserving edit. Clinical Transition Accuracy (CTA) measures the expected post-intervention response and does not condition on correctness of the original prediction. Matched Response Contrast (MRC) summarizes their separation. MRC is descriptive and is not interpreted as a causal effect.

3.3 AUTOMATED INTERVENTION EVIDENCE

Sham response stability alone does not characterize the intervention. We therefore compute six automated evidence terms using a pretrained chest-radiograph classifier and image similarity measures. Let p^o, p^s, p^c denote classifier outputs for the original, sham, and clinical images; e denote the target finding; and $\Delta_{i,g} = p^c_{i,e} - p^o_{i,e}$. The evidence terms are

$$v_{i,g}^{\mathrm{target}} = \mathrm{clip}\big(0.5 + (2a_i^{\mathrm{cf}} - 1)\Delta_{i,g}, 0, 1\big), \tag{6}$$

$$v_{i,g}^{\mathrm{pres}} = \mathrm{clip}\left(1 - \frac{1}{|\mathcal{K}_{-e}|}\sum_{k\in\mathcal{K}_{-e}} |p^c_{i,k} - p^o_{i,k}|, 0, 1\right), \tag{7}$$

$$v_{i,g}^{\mathrm{struct}} = \mathrm{clip}\Big(\mathrm{SSIM}(I_i, I_i^{\mathrm{cf},g}), 0, 1\Big), \tag{8}$$

$$v_{i,g}^{\mathrm{img}} = \mathrm{clip}\Big(1 - 2\ \mathrm{MAE}(I_i, I_i^{\mathrm{cf},g}), 0, 1\Big), \tag{9}$$

$$v_{i,g}^{\mathrm{sham}} = \frac{1}{2}\Big[\mathrm{SSIM}(I_i, I_i^{\mathrm{sham},g}) + \mathrm{clip}(1 - |p^s_{i,e} - p^o_{i,e}|, 0, 1)\Big], \tag{10}$$

$$v_i^{\mathrm{gen}} = \frac{1}{2}\Big[\mathrm{SSIM}(I_i^{\mathrm{cf},g_1}, I_i^{\mathrm{cf},g_2}) + \mathrm{clip}(1 - |p^{c,g_1}_{i,e} - p^{c,g_2}_{i,e}|, 0, 1)\Big]. \tag{11}$$

These terms represent target modification, non-target preservation, structural preservation, low-level image change, sham preservation, and cross-generator consistency, respectively. They are automated proxies, not expert assessments of clinical validity.

We aggregate them using the weighted geometric score

$$S_{\mathrm{IV}}(i,g) = \exp\left(\sum_{k=1}^{6} w_k \log\Big[\max(v_{i,g}^{(k)}, 10^{-6})\Big]\right), \tag{12}$$

with

$$\mathbf{w} = (0.25, 0.15, 0.15, 0.10, 0.20, 0.15).$$

Available weights are renormalized if a component is missing. S_{IV} is an automated composite score, not a calibrated probability of clinical validity. The weights are fixed design choices, not learned from clinician-labelled validity data.

3.4 THREE-WAY INTERVENTION DECISION

Prediction uncertainty and intervention uncertainty are distinct. We apply three-way decision theory Yao (2010; 2011) to the counterfactual evidence:

$$\delta(i,g) = \begin{cases} \mathrm{Accept}, & S_{\mathrm{IV}}(i,g) \geq \alpha, \\ \mathrm{Defer}, & \beta < S_{\mathrm{IV}}(i,g) < \alpha, \\ \mathrm{Reject}, & S_{\mathrm{IV}}(i,g) \leq \beta. \end{cases} \tag{13}$$

Table 1: **Sham-controlled VLM responses.** MRC is the descriptive contrast CTA−SFR.

Generator	Orig.	CTA	SFR	MRC
RadEdit	83.2%	67.0%	11.6%	55.4%
InstructX2X	83.2%	73.4%	44.8%	28.6%

We use $(\alpha, \beta) = (0.75, 0.45)$ as the reference operating point. These thresholds are fixed for the primary analysis but are neither learned from the evaluation cases nor treated as universal or Bayes-optimal boundaries. To test dependence on this choice, we evaluate all valid pairs from

$$\alpha \in \{0.65, 0.70, 0.75, 0.80\}, \qquad \beta \in \{0.30, 0.40, 0.45, 0.50\}, \qquad \beta < \alpha.$$

Thus, Accept, Defer, and Reject represent decisions under a specified operating policy rather than intrinsic labels of intervention validity.

3.5 Analysis

We compare paired S_{IV} values across generators using the Wilcoxon signed-rank test and report results separately by generator. We additionally examine threshold sensitivity and diagnostic score variants that remove sham or cross-generator evidence, together with binary and unfiltered evidence rules. These are diagnostic comparisons; no VLM or counterfactual generator is retrained.

4 Results and Discussion

4.1 Experimental Setup

We evaluate ValidCF on 500 patient-distinct NIH ChestX-ray14 cases, balanced between 250 effusion-positive and 250 effusion-negative patients. Pleural effusion is edited in both directions: Present→Absent and Absent→Present. We use RadEdit Pérez-García et al. (2024) and InstructX2X Min et al. (2025), producing 1,000 generator–case pairs. CheXagent is the downstream medical VLM. The experiment contains 3,000 VLM queries with no inference failures and 2,500 automated evidence evaluations: 500 original, 1,000 sham, and 1,000 clinical images.

4.2 Sham Controls Change the Generator Comparison

The VLM achieves 83.2% accuracy on the original images. InstructX2X achieves higher Clinical Transition Accuracy (CTA) than RadEdit (73.4% vs. 67.0%). However, its Sham Flip Rate (SFR) is also substantially higher (44.8% vs. 11.6%). The corresponding Matched Response Contrasts (MRC) are therefore 28.6% and 55.4%, respectively (Table 1).

The generator favored by CTA is therefore not the generator with the stronger separation between expected clinical response and sham instability (Fig. 2). This does not establish that the CTA difference is caused by editing artifacts, because sham and clinical edits need not have identical nuisance distributions. It instead shows that target-transition accuracy alone gives an incomplete account of the stability of counterfactual evidence.

4.3 Intervention Evidence and Three-Way Decisions

RadEdit obtains a higher mean intervention-evidence score S_{IV} than InstructX2X (0.792 vs. 0.704). The paired difference is significant under a Wilcoxon signed-rank test ($W = 8430$, $p = 4.286 \times 10^{-63}$). InstructX2X has stronger target-modification evidence (0.759 vs. 0.682), whereas RadEdit has stronger sham-preservation evidence (0.893 vs. 0.662). Thus, stronger target modification does not necessarily imply stronger target-preserving stability.

At the reference operating point $(\alpha, \beta) = (0.75, 0.45)$, 50.0% of interventions are assigned to Accept, 49.4% to Defer, and 0.6% to Reject. RadEdit assigns 76.6%/22.4%/1.0% to Accept/Defer/Reject, compared with 23.4%/76.4%/0.2% for InstructX2X (Table 2).

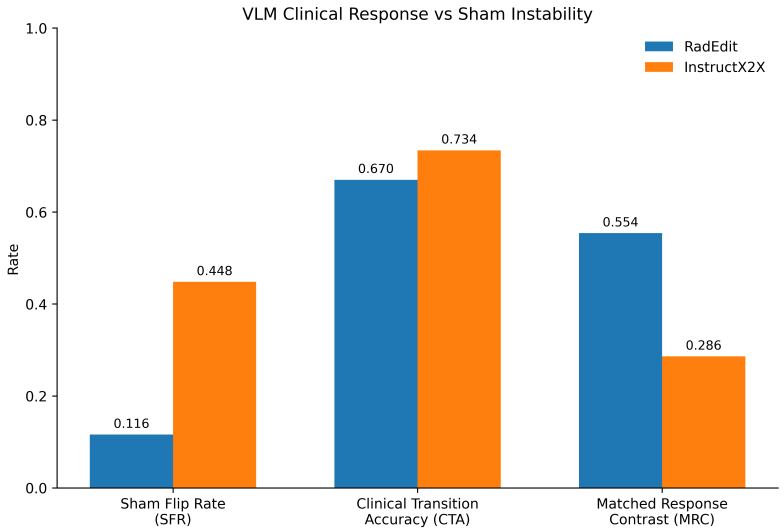

Figure 2: **Sham-controlled response comparison.** InstructX2X achieves higher CTA but also substantially higher sham instability, producing a lower MRC than RadEdit.

Table 2: **Three-way intervention decisions** at $(\alpha, \beta) = (0.75, 0.45)$.

Generator	S_{IV}	Accept	Defer	Reject
RadEdit	0.792	76.6%	22.4%	1.0%
InstructX2X	0.704	23.4%	76.4%	0.2%
Overall	–	50.0%	49.4%	0.6%

The 49.4% Defer rate does not mean that these interventions are clinically invalid. It means that their automated evidence lies between the two decision boundaries under the reference operating policy. This intermediate region avoids forcing ambiguous intervention evidence into a binary valid/invalid decision.

4.4 Threshold Sensitivity

We evaluate $\alpha \in \{0.65, 0.70, 0.75, 0.80\}$ and $\beta \in \{0.30, 0.40, 0.45, 0.50\}$ for all $\beta < \alpha$. As expected, the three-way allocation changes with the operating policy. For example, at $\beta = 0.45$, increasing α from 0.65 to 0.80 reduces the Accept rate from 88.8% to 27.4% and increases the Defer rate from 10.6% to 72.0%, while the Reject rate remains 0.6%.

This analysis shows that the three-way allocation should be interpreted relative to the selected operating policy. The continuous intervention evidence remains unchanged; the thresholds determine how that evidence is translated into Accept, Defer, and Reject decisions. Thus, $(\alpha, \beta) = (0.75, 0.45)$ is a reference operating point rather than an intrinsically optimal boundary.

4.5 Discussion and Limitations

The central finding is not that one generator is universally superior. Rather, CTA and sham-controlled evaluation support different interpretations: InstructX2X produces more expected clinical transitions but substantially greater response instability under target-preserving edits. This separates *prediction uncertainty* from *intervention uncertainty*. ValidCF evaluates the latter. Its Accept, Defer, and Reject decisions apply to synthetic counterfactual evidence, not patient diagnoses. The sham control is also a negative control, not the unobserved true patient counterfactual, and MRC is a descriptive contrast rather than a causal effect.

Several limitations remain. S_{IV} uses automated proxies; SSIM and pixel differences do not replace expert clinical assessment, and its aggregation weights are fixed rather than learned from clinician-labelled validity data. Likewise, the three-way thresholds are operating choices rather than calibrated clinical boundaries, although we explicitly evaluate their sensitivity. Finally, the study considers one chest-radiograph dataset, one target finding, one medical VLM, and two counterfactual generators. Broader clinical validation is therefore needed. Within this scope, the results show that counterfactual evaluation should examine the intervention producing the evidence, not only the model response that follows it.

5 Conclusion

We introduced ValidCF to evaluate synthetic counterfactual evidence before using it to interpret medical VLM behavior. ValidCF combines target-preserving sham controls with intervention evidence and three-way Accept, Defer, and Reject decisions. Across 500 patient-distinct chest radiographs, InstructX2X achieved higher clinical transition accuracy than RadEdit (73.4% vs. 67.0%), but also substantially higher sham response instability (44.8% vs. 11.6%). Consequently, RadEdit achieved a higher matched response contrast (55.4% vs. 28.6%). These results show that target-transition accuracy alone can give an incomplete account of counterfactual evidence sensitivity. Synthetic interventions should therefore be evaluated before their induced responses are interpreted as evidence of clinical sensitivity.

References

Chun-Hao Chang, George Alexandru Adam, and Anna Goldenberg. Towards robust classification model by counterfactual and invariant data generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15212–15221, 2021.

Lauren H. Cooke, Matthias Jung, Jan M. Brendel, Nora M. Kerkovits, Borek Foldyna, Michael T. Lu, and Vineet K. Raghu. Roentmod: A synthetic chest x-ray modification model to identify and correct image interpretation model shortcuts. *npj Digital Medicine*, 9:324, 2026.

Nima Fathi, Amar Kumar, Brennan Nichyporuk, Mohammad Havaei, and Tal Arbel. Decodex: Confounder detector guidance for improved diffusion-based counterfactual explanations. In *Proceedings of The 7th International Conference on Medical Imaging with Deep Learning*, volume 250 of *Proceedings of Machine Learning Research*, pp. 439–451. PMLR, 2024.

Hanqi Jiang, Junhao Chen, Mingyu Kang, Hyeokjae Kwon, Yi Pan, Lifeng Chen, Weihang You, Haozhen Gong, Ruiyu Yan, Jinglei Lv, Lin Zhao, Hui Ren, Quanzheng Li, Tianming Liu, and Xiang Li. Medvigil: Evaluating trustworthy medical vlms under broken visual evidence. *arXiv preprint arXiv:2605.07919*, 2026.

Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388, 2010.

Hyungi Min, Taeseung You, Hangyeul Lee, Yeongjae Cho, and Sungzoon Cho. Instructx2x: An interpretable local editing model for counterfactual medical image generation. In *Medical Image Computing and Computer Assisted Intervention*, pp. 279–288, 2025.

Dung Nguyen, Minh Khoi Ho, Huy Ta, Thanh Tam Nguyen, Qi Chen, Kumar Rav, Quy Duong Dang, Satwik Ramchandre, Son Lam Phung, Zhibin Liao, Minh-Son To, Johan Verjans, Phi Le Nguyen, and Vu Minh Hieu Phan. Localizing before answering: A benchmark for grounded medical visual question answering. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 7670–7678, 2025.

Yushu Pan and Elias Bareinboim. Counterfactual image editing. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39087–39101. PMLR, 2024.

Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Fernando Pérez-García, Sam Bond-Taylor, Pedro P. Sanchez, Boris van Breugel, Daniel C. Castro, Harshita Sharma, Valentina Salvatelli, Maria Teodora A. Wetscherek, Hannah C. M. Richardson, Matthew Lungren, Aditya Nori, Javier Alvarez-Valle, Ozan Oktay, and Maximilian Ilse. Radedit: Stress-testing biomedical vision models via diffusion image editing. In *European Conference on Computer Vision*, 2024.

Mélanie Roschewitz, Fabio De Sousa Ribeiro, Tian Xia, Galvin Khara, and Ben Glocker. Robust image representations with counterfactual contrastive learning. *Medical Image Analysis*, 105: 103668, 2025.

Emma A. M. Stanley, Vibujithan Vigneshwaran, Erik Y. Ohara, Finn G. Vamosi, Nils D. Forkert, and Matthias Wilms. Synthetic ground truth counterfactuals for comprehensive evaluation of causal generative models in medical imaging. In *Medical Image Computing and Computer Assisted Intervention*, pp. 541–550, 2025.

Guangzhi Xiong, Qiao Jin, Sanchit Sinha, Zhiyong Lu, and Aidong Zhang. Rethinking visual attribution for chest x-ray reasoning in large vision language models. *arXiv preprint arXiv:2605.20158*, 2026.

Yiyu Yao. Three-way decisions with probabilistic rough sets. *Information Sciences*, 180(3):341–353, 2010.

Yiyu Yao. The superiority of three-way decisions in probabilistic rough set models. *Information Sciences*, 181(6):1080–1096, 2011.

Jusheng Zhang, Kaitong Cai, Yijia Fan, Jian Wang, and Keze Wang. Cf-vlm: Counterfactual vision-language fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Kaifan Zhang, Lihuo He, Yuqi Ji, Junjie Ke, Lukun Wu, Tianhao You, and Xinbo Gao. Eeg-editbench: Probing visual information in eeg-image retrieval models with controlled image edits. *arXiv preprint arXiv:2607.27857*, 2026.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. In *Advances in Neural Information Processing Systems*, 2023.