Counterfactual Activation Patching of Pleural-Effusion Boundary Failures in MedSAM

Indira Duisembayeva[1], Ziyue Zhang[1], Muhammad Bilal[2], and Yutong Xie[1]

[1] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates

[2] Birmingham City University, Birmingham, United Kingdom

Abstract. Medical segmentation foundation models can fail when pathology obscures anatomical boundaries, but output masks alone do not explain which coarse encoder sites and spatial regions support the failure or its correction. We propose *counterfactual activation patching*, an interventional framework that pairs each diseased image with a patient-matched pseudo-healthy counterfactual and replaces selected MedSAM image-encoder activations between the two forward passes. We study lung segmentation in 200 pleural-effusion radiographs from 59 NIH ChestX-ray14 subjects using automatically generated CheXmask reference masks (silver labels). Counterfactual images improve MedSAM segmentation substantially: CF-Seg increases Dice from 0.726 to 0.850 and reduces HD95 from 123.5 to 80.0 pixels, with a second diffusion generator showing smaller but consistent gains. Activation-patching controls show that patient-matched patches reproduce counterfactual behavior, whereas non-matched donors perform substantially worse. Whole-map neck patching is treated as a pipeline sanity check; full-context patches give the strongest regional signal, but equal-area controls show that this is not a size-controlled context-pixel localization claim. These results position patient-matched counterfactuals as a controlled probe for analyzing and internally monitoring boundary failures in promptable medical segmentation, while expert annotations and external cohorts remain necessary before clinical-deployment claims. Our `code and aggregate results` are publicly available.

Keywords: Interventional analysis · Medical foundation models · MedSAM · Activation patching · Counterfactual imaging · Chest radiography

1 Introduction

Promptable segmentation foundation models provide a common interface for segmenting medical structures across datasets and modalities. The Segment Anything Model (SAM) introduced prompt-conditioned segmentation with an image encoder, prompt encoder, and mask decoder [1]; MedSAM adapts this design to large-scale medical image-mask data [2]. Complementary work uses domain-routed conditional residual modulation to adapt frozen vision transformers across

medical and natural image domains [3]. However, direct SAM transfer to medical images remains uneven because medical targets often have low contrast, ambiguous extent, prompt sensitivity, and domain-specific texture statistics [4,5,6]. Pleural effusion in chest radiography is a representative weak-boundary setting: fluid can obscure the lower lung contour, causing under-segmentation or opacity-aligned masks. Related chest-radiograph restoration work also emphasizes structural and edge preservation, using multi-wavelet inputs to denoise simulated low-dose images [7]. Dice, the boundary F1 score, and 95th-percentile Hausdorff distance (HD95) quantify segmentation error, but they do not identify whether a failure arises from local edge evidence, global shape context, prompt sensitivity, or later decoding.

Counterfactual medical imaging offers a controlled handle on this question. CF-Seg is closest to our setting because it generates pseudo-healthy chest radiographs from pleural-effusion cases and uses them to improve segmentation [8]. Other work uses generated radiographs to ask how an image or model prediction would change if pathology were absent, including chest-X-ray counterfactual explanations and diffusion-based medical anomaly editing [9,10,11]. Beyond segmentation, *MedObvious* shows that fluent medical vision-language-model outputs can coexist with failures on basic input-validity checks, reinforcing the need to evaluate reliability beyond apparently plausible outputs [12]. We ask a more mechanistic question: if a patient-matched pseudo-healthy counterfactual restores the lung boundary, which coarse MedSAM encoder sites and spatial regions carry the restoration under intervention?

We propose *counterfactual activation patching*: MedSAM is run on a diseased image and its patient-matched pseudo-healthy counterfactual with the same box prompt, then selected image-encoder activations are replaced across the paired runs. Matched, reverse, non-matched, and spatial patches test restoration, directionality, case-matched dependence, and localization. This paired-run logic adapts activation-patching methods that test whether internal states restore or suppress a behavior [13,14,15], framed as a coarse interventional analysis informed by causal mediation [16]. Our interventions are spatial/site-level probes rather than fine-grained circuit discovery over individual attention heads, channels, or learned features.

Figure 1 summarizes the paired inputs, interventions, controls, and downstream readouts.

Contributions.

- We introduce counterfactual activation patching as a coarse interventional probe for boundary failures in promptable medical segmentation.
- On 200 NIH ChestX-ray14 pleural-effusion radiographs from 59 subjects with CheXmask silver labels, we show that patient-matched patches reproduce counterfactual behavior, while reverse and non-matched controls support directionality and patient specificity.
- Full-context neck patches give the strongest regional Boundary F1 signal, but equal-area controls show the effect reflects broad late-token intervention

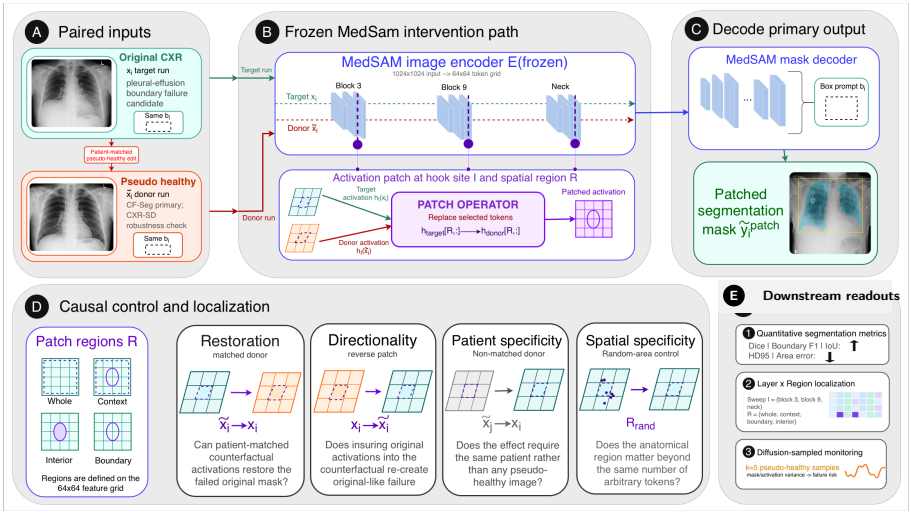

Fig. 1. Counterfactual activation patching workflow. (A) Each original CXR x_i is paired with a patient-matched pseudo-healthy edit $\tilde{x}_i$ from CF-Seg (primary) or CXR-SD (robustness), using the same box prompt b_i. (B) In frozen MedSAM, target and donor runs are intercepted at block 3, block 9, or the neck; region R tokens replace all target channels. (C) The decoder produces the primary patched mask. (D) Causal controls test restoration, directionality, patient specificity, and spatial specificity over whole, context, interior, and boundary regions on the 64×64 grid. (E) Readouts summarize segmentation metrics, layer-region localization, and diffusion-sampled monitoring over $K = 5$ samples.

rather than size-controlled context-pixel localization; diffusion-sampled variance provides an internal failure-ranking signal.

2 Methods

2.1 Data, Counterfactuals, and MedSAM Inference

We use 200 frontal NIH ChestX-ray14 radiographs labeled with pleural effusion from 59 subjects [17]. After filtering eligible AP/PA radiographs, records are sorted by subject and image identifiers and the first 200 are retained. Lung reference masks are obtained from CheXmask and treated as silver labels rather than expert annotations [18]. Left and right lungs are combined into one target. A bounding box is derived from this target with a 12-pixel margin and reused for original and counterfactual variants so that changes reflect image content and activations rather than prompt variation. The primary counterfactual source is CF-Seg pseudo-healthy generation for all 200 radiographs [8]. The secondary source, CXR-SD (CXR Stable Diffusion), uses the fixed `IrohXu/stable-diffusion-mimic-cxr-v0.1` checkpoint [19,20] in image-to-image mode with a no-findings prompt, a pleural-effusion/artifact

negative prompt, strength 0.35, guidance 7.5, and 40 steps. The generator operates at 512×512 pixels and outputs are bicubically resized to the original radiograph dimensions. Unless labeled otherwise, detailed tables report CF-Seg; matched/reverse/non-matched patching, layer-region localization, and boundary-width robustness are repeated on CXR-SD for the same 200 radiographs. Prompt-jitter robustness is evaluated with CF-Seg. Diffusion-sampled monitoring uses CXR-SD only because it requires multiple stochastic samples. Generated files are checked for existence, dimensions, and nonblank content, and representative images are visually inspected. We use the MedSAM ViT-B checkpoint; masks are generated from each original, CF-Seg, and CXR-SD image with the same box prompt and thresholded at 0.5.

2.2 Counterfactual Activation Patching

Let x_i be an original pleural-effusion image, $\tilde{x}_i$ its patient-matched pseudo-healthy counterfactual, m_i the silver lung mask, and b_i the fixed box prompt. MedSAM segmentation is

$$\hat{y}_i = D(E(x_i), b_i), \tag{1}$$

where E is the image encoder and D the prompt-conditioned decoder. At encoder site ℓ, patching replaces all channels within spatial region R of the target activation with the donor activation:

$$h_\ell^{\text{patch}}(u, v, c) = \begin{cases} h_\ell(x_{\text{donor}})(u, v, c), & (u, v) \in R, \\ h_\ell(x_{\text{target}})(u, v, c), & (u, v) \notin R. \end{cases} \tag{2}$$

Here (u, v) indexes a token location on the 64×64 feature grid and c indexes the activation channel. The patched activation is passed through the remaining computation and decoded with the same box prompt. PyTorch forward hooks patch image-encoder blocks 3 and 9 and the encoder neck. Block patches replace the full post-block tensor after the transformer residual update; neck patches replace the post-convolution/normalization output immediately before mask decoding. Inputs use a 1024 by 1024 image and a 64 by 64 feature grid; selected spatial tokens replace all channels.

For each generator, matched counterfactual-to-original patching uses donor $\tilde{x}_i$ and target x_i; reverse patching swaps this direction; non-matched patching uses donor $\tilde{x}_j$ from a different subject. Regions include the whole map, boundary bands around the CheXmask contour, lung interior, context outside the lung, and matched-area random controls. Boundary bands are dilation-minus-erosion masks at 16, 32, or 64 pixels, nearest-neighbor resized to the feature grid. In a 50-radiograph reference subset, the 32-pixel band covers 750.2 ± 82.7 of 4096 tokens (18.3%), while context covers 3133.6 ± 173.7 tokens (76.5%). To address this size imbalance, we add equal-area context and random controls that patch the same number of tokens as the 32-pixel boundary band, using three seeded repeats averaged per radiograph. CF-Seg prompt robustness uses 5%, 10%, and 15% deterministic box jitter; boundary-width sweeps are run for both generators.

2.3 Diffusion-Sampled Failure Monitoring

To test whether counterfactual variability flags poor-performing segmentations under the silver-label metrics, we generate five CXR-SD pseudo-healthy samples per radiograph using deterministic base seeds 17, 23, 31, 43, and 59, and run MedSAM with the same box prompt. We compute mask-variance features, neck-activation standard-deviation features, area variance, and pairwise Dice. Failures are median splits of original MedSAM performance: low Dice, low Boundary F1, or high HD95. Class-balanced logistic-regression probes use 20 repeats of subject-grouped stratified 5-fold cross-validation, with identical folds across feature sets. Confidence intervals use 2,000 subject-cluster bootstrap resamples. As an internal hold-out check, 100 subject-grouped 70/30 splits estimate the failure threshold and fit the probe on training radiographs only, then evaluate area under the receiver-operating-characteristic curve (AUROC) on held-out subjects.

2.4 Metrics and Statistical Testing

Predicted masks are evaluated against CheXmask using Dice, intersection over union (IoU), Boundary F1 with a 3-pixel tolerance, HD95 in pixels, and signed relative area error:

$$\Delta A_i = \frac{|\hat{m}_i| - |m_i|}{|m_i| + \epsilon}, \tag{3}$$

where positive values indicate over-segmentation. This metric set combines overlap, boundary, distance, and signed-size readouts [21]. We report image-level mean differences with 10,000 subject-cluster bootstrap confidence intervals. Two-sided Wilcoxon signed-rank tests are applied to the 59 subject-level mean differences for pre-specified mechanistic comparisons; unadjusted p values are interpreted alongside effect sizes. Because labels are silver, conclusions are framed as controlled mechanistic evidence rather than deployment-ready clinical validation.

3 Experiments and Results

3.1 Counterfactual Inputs Improve MedSAM Masks

Table 1 compares MedSAM predictions on original diseased images, CF-Seg pseudo-healthy counterfactuals, and the CXR-SD diffusion counterfactuals. Across 200 radiographs, CF-Seg improves Dice from 0.726 to 0.850, Boundary F1 from 0.156 to 0.217, and HD95 from 123.5 to 80.0 pixels. CXR-SD also improves all three metrics, but less strongly (Dice 0.800, Boundary F1 0.172, HD95 103.9), so we treat it as an independent generator-diversity check rather than a stronger segmentation generator.

Table 1. MedSAM segmentation on original and counterfactual images. Higher is better except for HD95; ΔA is signed relative area error.

Condition	n	Dice	IoU	Boundary F1	HD95	ΔA
Original, main cohort	200	0.726	0.587	0.156	123.5	0.324
CF-Seg, main cohort	200	0.850	0.754	0.217	80.0	0.178
CXR-SD, main cohort	200	0.800	0.679	0.172	103.9	0.284

The paired effect is broad: CF-Seg improves Dice in 92.0% of cases, Boundary F1 in 74.0%, and HD95 in 80.5%; CXR-SD improves them in 78.5%, 58.5%, and 65.0%. The remaining failures motivate the controls below: if counterfactual images simply made every case easier, non-matched or arbitrary patches should also work, but they do not.

3.2 Whole-Map Patching Is a Pipeline Sanity Check

Table 2 shows the 200-radiograph CF-Seg encoder-neck intervention controls: matched whole-map patching recovers counterfactual behavior, reverse patching restores original-like behavior, and non-matched patching is substantially worse. Because whole-map neck patching replaces the final image embedding before the same box-conditioned decoder, this result is expected by construction and serves as a pipeline sanity check rather than fine-grained mechanistic evidence. The CXR-SD run shows the same direction, with matched whole-map Boundary F1 0.172 versus non-matched 0.055 and reverse 0.156 (Fig. 2, left), supporting a case-matched effect rather than generic healthy-feature injection.

Table 2. Main CF-Seg activation-patching controls at the encoder neck on 200 radiographs.

Condition	Dice	IoU	Boundary F1	HD95	ΔA
Original	0.726	0.587	0.156	123.5	0.324
Matched whole	0.850	0.754	0.217	80.0	0.178
Reverse whole	0.726	0.587	0.156	123.5	0.324
Non-matched whole	0.662	0.507	0.057	128.3	0.177
Context	0.824	0.722	0.251	86.3	0.190
Boundary band	0.761	0.635	0.164	107.3	0.240
Interior	0.787	0.662	0.174	110.5	0.330

Under the original unequal-size region definitions, full-context patching gives the highest Boundary F1 (0.251), while boundary and interior patches are weaker. Because full context contains substantially more tokens, we treat this as a broad late-representation effect and evaluate size-matched controls below.

3.3 Layer and Region Controls Identify a Broad Context Effect

Figure 2 summarizes the scaled 200-radiograph encoder-neck controls. Extending the same matched counterfactual-to-original localization to block 3 and block 9 gives the same ordering under the original unequal-size definitions: full context remains stronger than the thin boundary band for both generators, and the largest regional Boundary F1 occurs at the neck (0.251 for CF-Seg and 0.196 for CXR-SD).

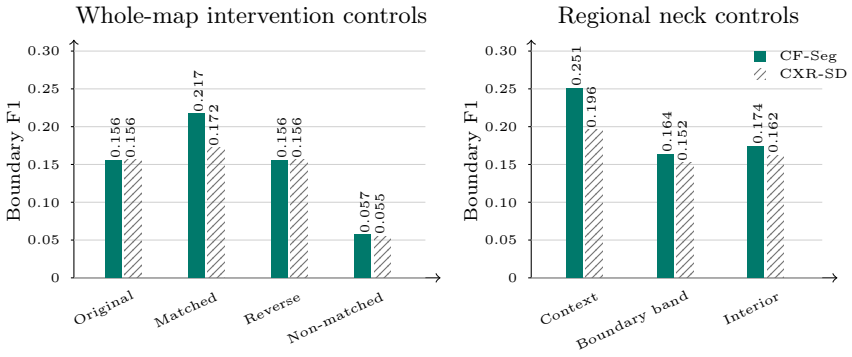

Fig. 2. Scaled 200-radiograph encoder-neck activation controls measured by Boundary F1. Left: whole-map sanity checks show matched patching improves over original and reverse behavior while non-matched donors fail. Right: regional controls show the strongest signal in context tokens under the tested, unequal-size region definitions, consistently across CF-Seg and CXR-SD. Hatched bars denote CXR-SD.

3.4 Boundary Errors Shift with Broader Context

Spatial controls test whether the corrective signal is concentrated in a thin anatomical boundary band. In the 200-radiograph CF-Seg neck run, full-context patching reaches Boundary F1 0.251, above boundary-band (0.164) and lung-interior patching (0.174); CXR-SD gives the same ordering with smaller effects (0.196, 0.152, 0.162; Fig. 2, right). Paired tests in Table 5 confirm full context $>$ boundary and full context $>$ interior for both sources. Equal-area controls change the interpretation (Table 3): context sampled to the same token count as the boundary band falls below the boundary band for CF-Seg and does not exceed it for CXR-SD. Thus, the robust finding is a broad full-context effect in late MedSAM representations, not proof that context pixels alone are the localized cause of repair. Because late MedSAM tokens integrate nonlocal information, a token's spatial coordinate also does not establish that the relevant evidence originated from the corresponding input pixels. Table 4 adds stress tests: for CF-Seg, matched whole-map patching remains stronger than non-matched patching under 5–15% prompt jitter. For both generators, full context remains stronger than

boundary under 16-, 32-, and 64-pixel width sweeps. Severe jitter can narrow or flip the context/boundary ordering, so the safest interpretation is a case-matched intervention effect with coarse, size-sensitive regional structure under controlled prompting.

Table 3. Equal-area neck controls measured by Boundary F1. Context and random controls patch the same number of tokens as the 32-pixel boundary band; three seeded repeats are averaged per case.

Source	Full ctx.	Boundary	Eq. ctx.	Eq. rand.
CF-Seg	0.251	0.164	0.140	0.143
CXR-SD	0.196	0.152	0.146	0.147

Table 4. NIH-200 neck robustness by Boundary F1. Jitter rows are CF-Seg experiments and use non-matched whole-map donors as controls; width rows use context-size random regions as controls.

Test	Setting	Whole	Control	Context	Boundary
Jitter	5%	0.118	0.050	0.132	0.110
Jitter	10%	0.061	0.038	0.068	0.082
Jitter	15%	0.040	0.032	0.048	0.064
CF-Seg width	16/32/64 px	0.217	0.182	0.251	0.156/0.164/0.185
CXR-SD width	16/32/64 px	0.172	0.158	0.196	0.151/0.152/0.157

Table 5. Paired statistics over 200 radiographs from 59 subjects; differences are candidate minus baseline. Confidence intervals resample subjects, and Wilcoxon tests use subject-level mean differences.

Comparison	Source	Metric	Mean diff.	95% bootstrap CI	p
Counterfactual vs. original	CF-Seg	Dice	+0.125	[0.099, 0.161]	3.09×10^{-11}
	CF-Seg	Boundary F1	+0.061	[0.041, 0.083]	4.26×10^{-8}
	CF-Seg	HD95	-43.45	[-54.71, -33.51]	1.29×10^{-9}
Counterfactual vs. original	CXR-SD	Dice	+0.074	[0.049, 0.108]	7.31×10^{-10}
	CXR-SD	Boundary F1	+0.016	[0.007, 0.030]	1.01×10^{-4}
	CXR-SD	HD95	-19.55	[-28.48, -12.17]	6.37×10^{-6}
Matched vs. non-matched whole	CF-Seg	Boundary F1	+0.160	[0.124, 0.204]	2.39×10^{-11}
	CXR-SD	Boundary F1	+0.117	[0.093, 0.152]	2.39×10^{-11}
Full context vs. boundary	CF-Seg	Boundary F1	+0.087	[0.077, 0.102]	2.65×10^{-11}
	CXR-SD	Boundary F1	+0.045	[0.038, 0.053]	3.27×10^{-9}
Full context vs. interior	CF-Seg	Boundary F1	+0.077	[0.066, 0.093]	1.47×10^{-10}
	CXR-SD	Boundary F1	+0.034	[0.026, 0.045]	1.72×10^{-8}

3.5 Diffusion Variance Predicts Failure

Diffusion-sampled pseudo-healthy variability predicts original-image MedSAM failure within this cohort (Table 6): output plus diffusion variance reaches Dice-failure AUROC 0.873 [0.796, 0.928], and diffusion mask variance reaches Boundary F1-failure AUROC 0.796 [0.714, 0.854], outperforming output-only and prompt-jitter baselines. A stricter 100-repeat subject-grouped 70/30 hold-out check estimates the median-failure threshold and trains the probe on training radiographs only; mask variance remains strong for low Dice/Boundary F1 (mean AUROC 0.844/0.787) versus output shape (0.639/0.486). Because labels and thresholds still come from CheXmask on NIH-200, these AUROCs show internal risk ranking rather than independent clinical validation.

Table 6. Failure prediction from diffusion variance. Entries are AUROC [95% CI].

Feature set	Low Dice	Low Boundary F1	High HD95
Output shape	0.607 [0.506, 0.719]	0.433 [0.367, 0.526]	0.780 [0.721, 0.850]
Prompt-jitter variance	0.666 [0.570, 0.782]	0.600 [0.487, 0.690]	0.567 [0.474, 0.664]
Diffusion mask variance	0.858 [0.766, 0.921]	0.796 [0.714, 0.854]	0.754 [0.670, 0.837]
Diffusion activation variance	0.840 [0.723, 0.910]	0.767 [0.664, 0.831]	0.702 [0.555, 0.802]
Output + diffusion variance	0.873 [0.796, 0.928]	0.778 [0.702, 0.832]	0.817 [0.718, 0.915]

With $K = 2, 3, 5$, diffusion mask variance remains predictive for low Dice (AUROC 0.795/0.821/0.858) and low Boundary F1 (0.738/0.749/0.796). Figure 3 illustrates high-variance failure cases; it does not constitute validation against clinically adjudicated failure labels.

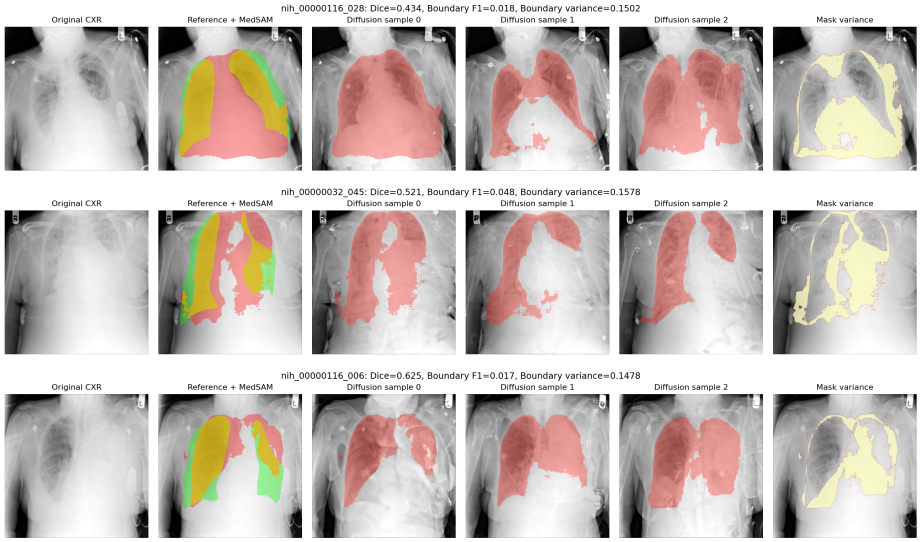

Fig. 3. Diffusion-sampled uncertainty in high-variance failure cases; variance concentrates around unstable context and boundary regions.

4 Discussion

Case-matched encoder interventions can shift MedSAM pleural-effusion boundary failures. The matched, reverse, and non-matched controls establish the intended directionality and patient-matched dependence of the intervention. Whole-map neck patching is expected to reproduce donor-like behavior because it replaces the final image embedding before the fixed prompt-conditioned decoder; it is therefore a pipeline sanity check rather than a fine-grained interpretability result. Under unequal-size regions, full-context neck patches produce the strongest signal, but equal-area controls show that this effect does not localize repair to context pixels. Together with the nonlocal receptive fields of late encoder tokens, the evidence supports a broad, size-sensitive late-representation effect. Diffusion-sampled variance additionally ranks likely failure cases within NIH-200, offering a monitoring use beyond the intervention analysis.

5 Limitations and Future Work

Several limitations bound these conclusions. First, evaluation uses CheXmask silver masks and reference-derived box prompts rather than expert pleural-effusion boundary annotations and clinically realistic prompts; improved agreement with these masks does not establish anatomical superiority. Second, pseudo-healthy counterfactuals can contain residual pathology, editing artifacts, or anatomy changes. File checks and visual inspection cannot verify pathology removal or

anatomical fidelity. Third, all evaluations use one deterministically selected subset of 200 radiographs from 59 NIH ChestX-ray14 subjects. Repeated observations from some subjects reduce cohort diversity, although subject-grouped splits and subject-clustered inference account for this dependence in monitoring and statistical testing. The second generator tests generator diversity on the same cohort, not external generalization. Fourth, patching at three encoder sites and broad token regions is a coarse intervention analysis: it does not identify individual heads, channels, features, or the input-pixel origin of information in late tokens. Finally, the monitoring targets are median splits of silver-label performance. Training-only threshold estimation and model fitting avoid direct use of held-out outcomes, and subject grouping prevents radiographs from the same subject appearing in both partitions; the analysis nevertheless remains internal validation rather than independently labeled failure detection.

Future work should evaluate the method with expert contours, clinically supplied prompts, and independent multi-site cohorts with prospectively defined failure labels. It should also assess anatomy-preserving counterfactual generators, additional promptable encoder–decoder models, and finer-grained feature-level interventions. These studies would test whether the observed case-matched intervention effect and variance-based risk ranking transfer beyond the present controlled setting.

6 Conclusion

We introduced counterfactual activation patching for analyzing pleural-effusion boundary failures in MedSAM. Patient-matched, reverse, and non-matched controls show that paired counterfactual activations can causally shift segmentation behavior, while equal-area controls appropriately limit the regional claim to broad late-representation effects. Diffusion-sampled variability provides an internal signal for ranking unreliable masks. The framework offers a controlled analysis tool for promptable medical segmentation, with expert and external validation needed before clinical interpretation or deployment.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4015–4026 (2023)
2. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 654 (2024). https://doi.org/10.1038/s41467-024-44824-z

3. Khan, U., Nawaz, U., Caputo, M., Bilal, M., Qadir, J., Khan, M.H.: Keep it frozen: Domain-routed conditional residual modulation for multi-domain vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21016–21025 (2026)
4. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: an experimental study. Medical Image Analysis **89**, 102918 (2023). https://doi.org/10.1016/j.media.2023.102918
5. Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., Liu, S., Chi, H., Hu, X., Yue, K., Li, L., Grau, V., Fan, D.P., Dong, F., Ni, D.: Segment anything model for medical images? Medical Image Analysis **92**, 103061 (2024). https://doi.org/10.1016/j.media.2023.103061
6. He, S., Bao, R., Li, J., Stout, J., Bjornerud, A., Grant, P.E., Ou, Y.: Computer-vision benchmark segment-anything model (SAM) in medical images: accuracy in 12 datasets. arXiv preprint arXiv:2304.09324 (2023)
7. Cohen, E.I., Khan, U., Wallace, B., Zandieh, A.R., Nezami, N., Filice, R.W., Field, D.H., Poulett, W., Ashraf, S., Bilal, M.: WE-UNet: A wavelet-enhanced U-Net framework for radiation dose reduction in chest radiography. Journal of Applied Clinical Medical Physics **27**(5), e70583 (2026). https://doi.org/10.1002/acm2.70583
8. Mehta, R., De Sousa Ribeiro, F., Xia, T., Roschewitz, M., Santhirasekaram, A., Marshall, D.C., Glocker, B.: CF-Seg: Counterfactuals meet segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. LNCS, vol. 15967, pp. 117–127. Springer (2025). https://doi.org/10.1007/978-3-032-04984-1_12
9. Atad, M., Dmytrenko, V., Li, Y., Zhang, X., Keicher, M., Kirschke, J., Wiestler, B., Khakzar, A., Navab, N.: CheXplaining in style: counterfactual explanations for chest X-rays using StyleGAN. arXiv preprint arXiv:2207.07553 (2022)
10. Cohen, J.P., Brooks, R., En, S., Zucker, E., Pareek, A., Lungren, M.P., Chaudhari, A.S.: The effect of counterfactuals on reading chest X-rays. arXiv preprint arXiv:2304.00487 (2023)
11. Wolleb, J., Bieder, F., Sandkühler, R., Cattin, P.C.: Diffusion models for medical anomaly detection. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, LNCS, vol. 13438, pp. 35–45. Springer (2022). https://doi.org/10.1007/978-3-031-16452-1_4
12. Khan, U., Nawaz, U., Teja, L.D.M.S.S., Saeed, N., Bilal, M., Xie, Y., Yaqub, M., Khan, M.H.: MedObvious: Exposing the medical Moravec's paradox in VLMs via clinical triage. arXiv preprint arXiv:2603.23501 (2026). https://doi.org/10.48550/arXiv.2603.23501
13. Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., Shieber, S.: Investigating gender bias in language models using causal mediation analysis. In: Advances in Neural Information Processing Systems, vol. 33, pp. 9573–9585 (2020)
14. Meng, K., Bau, D., Andonian, A., Belinkov, Y.: Locating and editing factual associations in GPT. In: Advances in Neural Information Processing Systems, vol. 35, pp. 17359–17372 (2022)
15. Zhang, F., Nanda, N.: Towards best practices of activation patching in language models: metrics and methods. In: International Conference on Learning Representations (ICLR) (2024). arXiv:2309.16042
16. Imai, K., Keele, L., Yamamoto, T.: Identification, inference and sensitivity analysis for causal mediation effects. Statistical Science **25**(1), 51–71 (2010). https://doi.org/10.1214/10-STS321

17. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2097–2106 (2017). https://doi.org/10.1109/CVPR.2017.369
18. Gaggion, N., Mosquera, C., Mansilla, L., Saidman, J.M., Aineseder, M., Milone, D.H., Ferrante, E.: CheXmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images. Scientific Data **11**, 511 (2024). https://doi.org/10.1038/s41597-024-03358-1
19. IrohXu: stable-diffusion-mimic-cxr-v0.1. Hugging Face model card. https://huggingface.co/IrohXu/stable-diffusion-mimic-cxr-v0.1 (accessed August 11, 2026)
20. Liang, K., Cao, X., Liao, K.-D., Gao, T., Ye, W., Chen, Z., Cao, J., Nama, T., Sun, J.: PIE: Simulating disease progression via progressive image editing. arXiv preprint arXiv:2309.11745 (2023). https://arxiv.org/abs/2309.11745
21. Maier-Hein, L., Reinke, A., Godau, P., et al.: Metrics reloaded: recommendations for image analysis validation. Nature Methods **21**, 195–212 (2024). https://doi.org/10.1038/s41592-023-02151-z