Counterfactual Lesion Erasure for Measuring Visual Grounding in Medical Vision-Language Explanation

Yahya Ely[1], Bakar Sidaty[2], and Hamadi Ahmed[3]

[1]Department of Applied Mathematics and Informatics, Higher Institute of Technological Education of Rosso, Route de Rosso, Rosso 16100, Mauritania
[2]Faculty of Legal and Economic Sciences, University of Aioun, Avenue de l'Indépendance, Aioun El Atrouss 51100, Mauritania
[3]Department of Information Systems and Management, École Normale Supérieure de Nouakchott, Avenue Moktar Ould Daddah, Nouakchott 11200, Mauritania

Abstract

Medical vision-language systems often produce convincing explanatory sentences, but fluency alone does not show that the explanation is visually grounded. A model may name the correct abnormality while basing its wording on surrounding context, memorized report patterns, or language priors rather than the actual image region. This paper develops an empirical counterfactual lesion-erasure framework for testing whether generated medical explanations depend on the image evidence they claim to describe. We assembled 5,760 radiographic and dermatologic image-question pairs with expert lesion masks, generated short explanatory answers using five multimodal models, and then created counterfactual images by selectively erasing, preserving, or relocating clinically relevant regions. A grounding-sensitive model should reduce abnormality-specific explanation after lesion erasure, preserve it when irrelevant background is erased, and avoid transferring the original explanation to anatomically inconsistent relocated evidence. We introduce three quantitative measures: counterfactual explanation collapse, region-preservation stability, and relocation contradiction rate. We also estimate a latent grounding coefficient using a structured matrix model linking mask overlap, explanation content, and answer correctness. Across all models, ordinary answer accuracy was 78.6%, but only 54.2% of correct explanations showed expected collapse after lesion erasure. The proposed grounding coefficient separated visually dependent explanations from language-driven explanations with an area under the receiver operating characteristic curve of 0.817 against expert labels. The results indicate that medical vision-language evaluation should directly perturb claimed evidence regions rather than infer grounding from correctness or attention maps alone.

1 Introduction

A medical vision-language model can answer a question correctly and still fail to use the part of the image that would justify the answer. This distinction is easy to miss because generated explanations often sound organized, cautious, and clinically familiar. A chest radiograph model may say that a left basilar opacity supports pneumonia, yet its answer may remain almost unchanged when that opacity is digitally removed. A dermatology model may mention irregular border and color variation, yet the same wording may appear after the lesion is replaced by healthy surrounding skin [1]. In these cases, the model's language imitates an explanation without demonstrating visual dependence. The present paper studies this failure through counterfactual lesion erasure [2].

The central problem is evidential attachment. When a generated sentence says that a finding is visible, the sentence implicitly points to a region in the image. If that region disappears, the sentence should change. If a different irrelevant region disappears, the sentence should remain mostly stable. If the region is moved to an implausible anatomical location, the answer should not continue as though the original image were unchanged. These expectations are simple for human interpretation, but they are not automatically satisfied by multimodal neural systems. Vision-language models can blend visual information with strong learned language patterns, and the balance between the two may differ across tasks, abnormalities, and prompting styles.

This paper does not treat explanation faithfulness as a philosophical property. It treats it as a measurable counterfactual relation. A generated explanation is visually grounded when targeted changes to the image evidence cause appropriate changes in the explanation. The model is not required to expose its internal attention weights

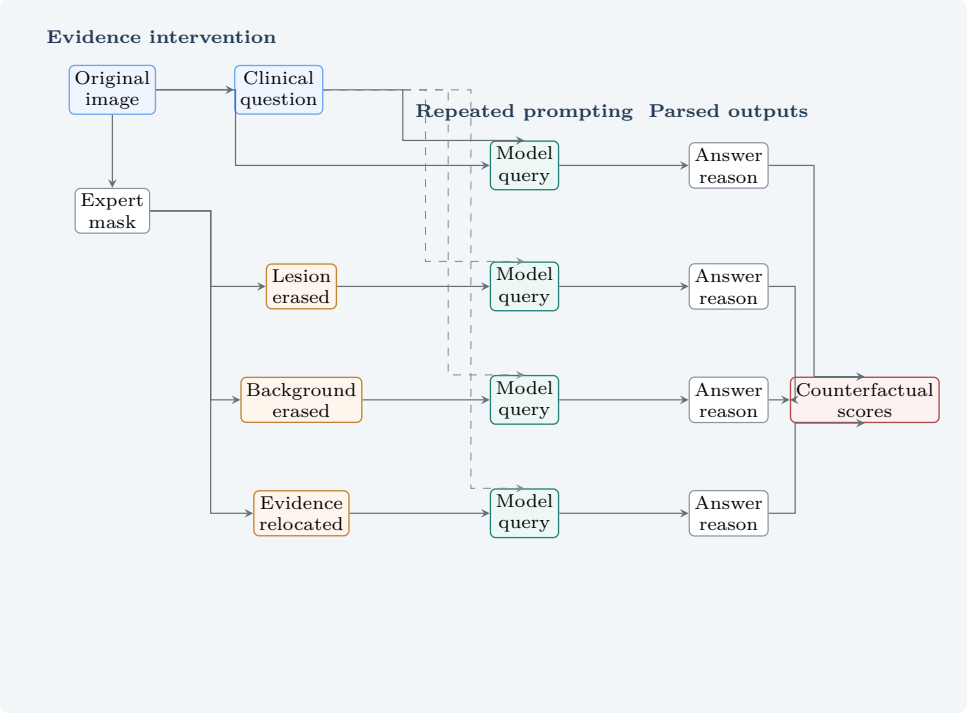

Figure 1: Counterfactual lesion-erasure evaluation pipeline. The same clinical question is paired with the original image and three controlled image variants, and the resulting short answer–reason pairs are compared to measure whether explanatory content follows the masked evidence.

[3]. It is not required to produce a full segmentation mask. It is required to behave consistently under controlled changes to image regions. The method therefore evaluates outputs rather than internal mechanisms. This is useful because many deployed systems expose only image input and text output.

The topic is separate from ordinary diagnostic accuracy. A model can be accurate because the abnormality is common in the dataset, because the question contains a clue, or because language priors point toward the reference answer. Accuracy alone cannot reveal whether the model relied on the lesion. Conversely, a model can be visually grounded but wrong if the image evidence is ambiguous or if the reference label is imperfect. Grounding and correctness overlap, but they are not the same. A complete evaluation should therefore measure both.

The topic is also separate from broad adversarial safety. Lesion erasure is not designed to trick the model with malicious instructions or hidden perturbations [4]. It is a controlled scientific intervention. The abnormal region is removed, preserved, or relocated while the prompt remains neutral. The goal is to observe whether explanatory language follows the image evidence. This makes counterfactual erasure closer to a causal probe than to an attack benchmark. The intervention asks what would happen to the explanation if the visual evidence were absent.

Medical imaging is especially suitable for this type of probe because many findings have spatial evidence. A pneumothorax has a pleural line and absent peripheral markings. A pleural effusion has dependent opacity and blunting. A fracture has cortical discontinuity or alignment abnormality. A skin lesion has border, color, asymmetry, scale, or ulceration. Expert masks can identify the region most relevant to the explanation. Once that region is known, one can ask whether the model's explanation responds to its removal. The same logic applies to other modalities, but this paper focuses on radiography and dermatology because they provide both localized visual evidence and interpretable textual explanations.

The paper introduces a benchmark with expert lesion masks and controlled counterfactual image transformations. Each original image is paired with a clinical question and a reference answer. The model first answers using the original image. Then the image is altered in three ways. In the lesion-erased image, the expert-marked region is replaced with surrounding-compatible texture or intensity. In the background-erased image, an irrelevant region of matched size is removed. In the relocated image, the lesion region is placed in an anatomically inconsistent or clinically irrelevant location while the original lesion site is erased. The same question is asked for each altered image. A grounded model should respond differently across these conditions.

The measurement framework has three primary outcomes. Counterfactual explanation collapse measures whether abnormality-specific language decreases after lesion erasure. Region-preservation stability measures whether the explanation remains stable when irrelevant background is erased. Relocation contradiction rate measures whether the model wrongly preserves the original explanation after lesion relocation. These outcomes together distinguish several behaviors. A model that changes after any image edit may be visually fragile rather than

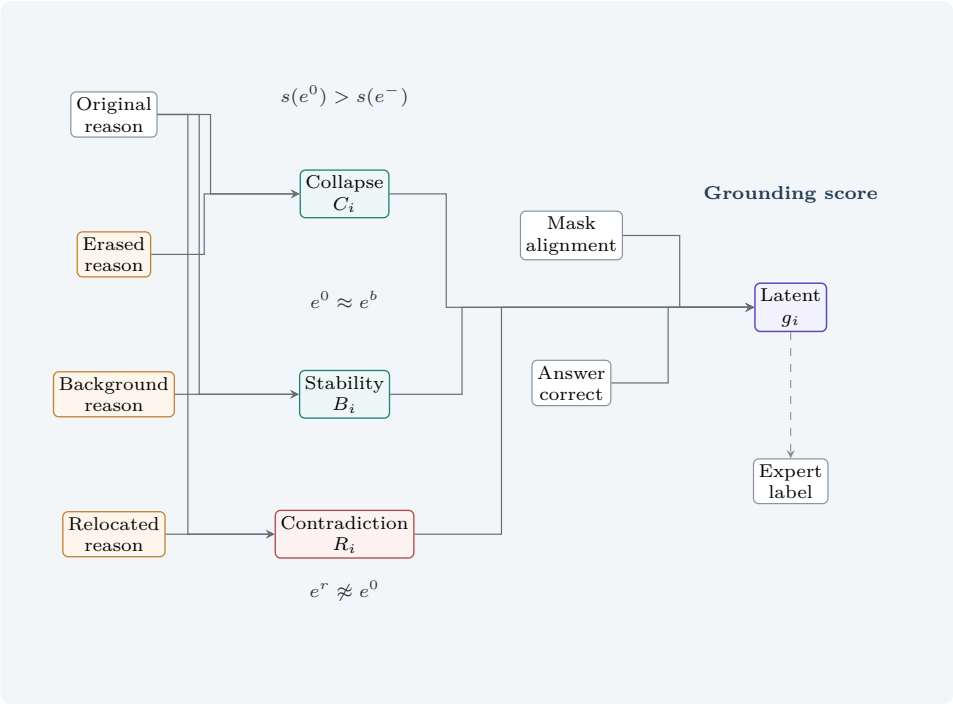

Figure 2: Metric fusion for estimating counterfactual grounding. Collapse, region-preservation stability, relocation contradiction, mask-language alignment, and answer correctness contribute complementary evidence for the latent coefficient used to identify visually dependent explanations.

grounded. A model that never changes may be language-driven. A model that changes only when the lesion region changes is closer to evidence-sensitive explanation.

A useful hint for designing evaluator aggregation comes from a different medical VLM testing setting. Sadanandan and Behzadan (2025) combined ratings from two independent judge models through z-score normalization before forming a composite vulnerability measure [5]. In the present work, that idea is not used to score vulnerability; instead, it motivates a narrower normalization choice in which two explanation graders with different response scales are standardized before their scores enter the latent grounding model. This citation is included for that specific evaluator-normalization detail and not as a template for the experimental design.

The empirical questions are direct. First, when medical vision-language models answer correctly, how often do their explanations respond to erasure of the visual evidence? Second, can counterfactual metrics distinguish explanation faithfulness from ordinary answer correctness? Third, does a latent grounding coefficient improve prediction of expert-labeled groundedness compared with attention overlap, answer accuracy, or explanation similarity alone? Fourth, do radiographic and dermatologic tasks show different patterns of counterfactual failure? The study answers these questions with generated outputs from five models, expert masks, human review, and statistical analysis.

The contributions are threefold. The first contribution is a counterfactual evaluation procedure that directly manipulates expert-marked visual evidence and examines the resulting explanatory text. The second contribution is a mathematical grounding model that links mask-sensitive visual change to explanation content through a latent coefficient. The third contribution is a set of empirical findings showing that correct answers frequently lack counterfactual evidence dependence. This last point is important because many model reports appear acceptable under ordinary answer scoring but fail when the claimed evidence is removed.

The rest of the paper proceeds as follows. The next section describes the dataset, image transformations, and explanation collection protocol. The following section defines the mathematical model for counterfactual grounding. The experimental section presents grading, statistical tests, and comparison baselines. The results section reports accuracy, collapse behavior, stability, relocation failures, model differences, and modality differences. The later sections interpret the findings, discuss limitations, and outline how counterfactual evidence testing may fit into future medical model evaluation.

2 Data Construction and Counterfactual Image Editing

The benchmark was assembled from two image families: localized radiographic findings and localized dermatologic lesions. The radiographic portion contained chest and extremity images because they include abnormalities with

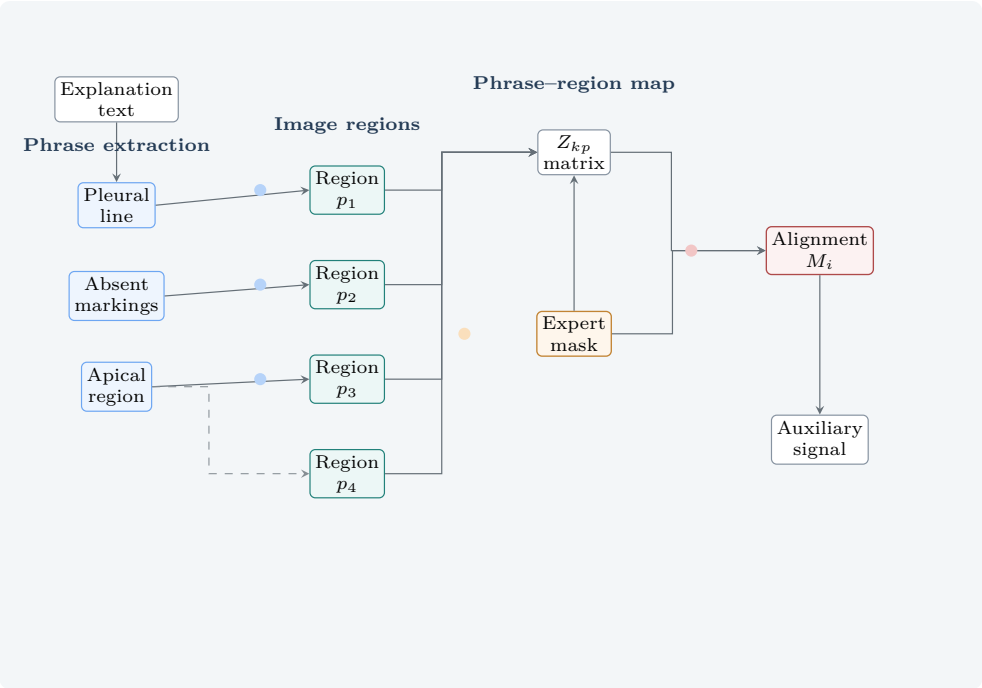

Figure 3: Structured mask-language alignment. Visual phrases from the explanation are matched against image regions, then aggregated through expert-mask overlap to form an auxiliary grounding signal that is later combined with counterfactual response metrics.

recognizable regional evidence. The dermatologic portion contained clinical skin photographs with lesion masks. The combined dataset contained 5,760 image-question pairs. Of these, 3,360 were radiographic and 2,400 were dermatologic. Each case had an expert mask identifying the visual region most relevant to the reference answer. The mask did not need to cover every related pixel; it needed to identify the principal evidence region whose removal should weaken the abnormality explanation.

The radiographic cases included pneumothorax, pleural effusion, focal opacity, fracture, dislocation, orthopedic hardware complication, and selected device-position abnormalities. The dermatologic cases included melanocytic lesion concern, inflammatory plaque, ulceration, pigment irregularity, border irregularity, and benign-appearing lesions used as controls. The dataset intentionally included both positive and negative cases. For negative cases, masks were assigned to visually salient but clinically irrelevant regions so that the erasure procedure could test whether the model invented explanations around irrelevant visual texture. Positive cases were oversampled because the main outcome requires a lesion region whose removal should change abnormality-specific explanation.

Each case was converted into a focused question. Radiographic questions asked about a particular abnormality or location, such as whether a pneumothorax was present, whether a fracture was visible, or whether a device tip was appropriately positioned. Dermatologic questions asked whether the lesion showed a particular visual property, such as irregular border, color variegation, ulceration, or inflammatory scale. Questions were written without giving away the answer. They did not mention the mask, the edited condition, or any expected abnormality beyond the clinical topic. For every case, a short reference answer and a reference explanation phrase were produced from expert annotation.

The expert mask collection followed a two-pass procedure. In the first pass, an annotator marked the smallest region that contained the main evidence for the reference finding. In the second pass, another annotator reviewed whether erasing that region would reasonably reduce the visual basis for the answer. Disagreements were resolved by consensus. For radiographs, masks were drawn on the original image resolution. For dermatology images, masks were drawn around the lesion and, when necessary, around subregions such as ulcerated focus or irregular border segment. The median mask area was 4.8% of image area for radiographs and 18.6% for dermatology images.

Three edited image variants were created for each original case. The first was the lesion-erased variant. The masked region was replaced using modality-specific inpainting. For radiographs, the replacement used local intensity interpolation guided by surrounding tissue patterns and an edge-smoothing constraint. For dermatology images, the replacement used patch-based synthesis from nearby nonlesional skin while preserving illumination [6]. The second was the background-erased variant. A region of matched area and approximate shape was selected from a clinically irrelevant part of the image, then erased with the same inpainting procedure. The third was the relocation variant. The lesion region was removed from its original position and inserted into a region where the same explanation would be anatomically inconsistent or clinically irrelevant.

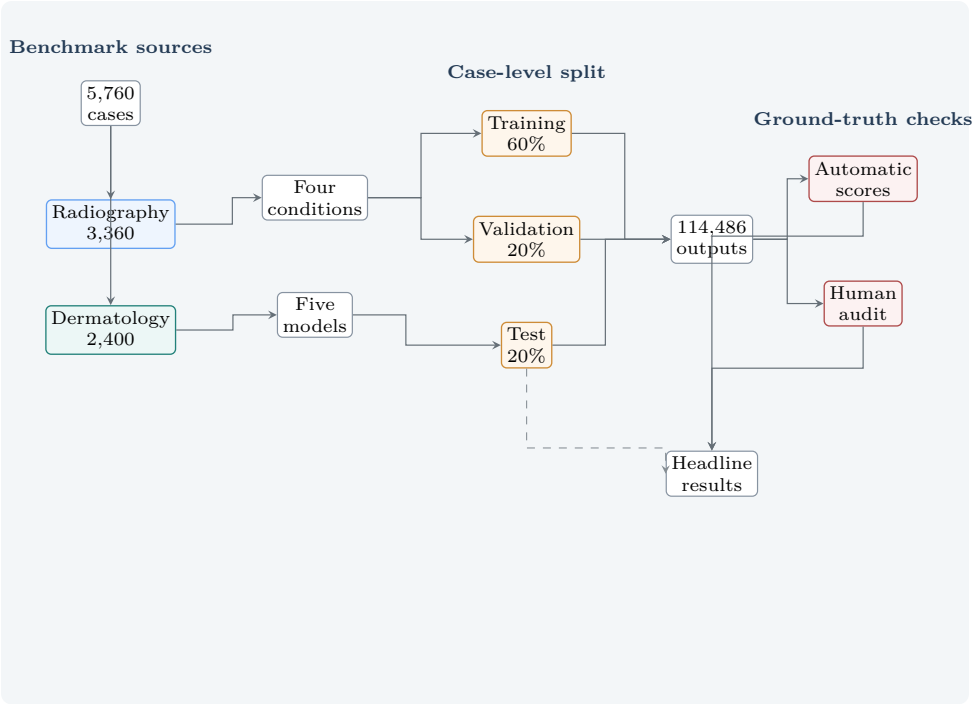

Figure 4: Experimental organization. Cases from radiographic and dermatologic sources are expanded across original and edited image conditions, queried across five systems, split by case group, and evaluated with automatic metrics plus blinded human audit.

Quality control was performed to avoid trivial artifacts. Human reviewers inspected a random sample of 1,200 edited images and rated whether the edit was visually obvious, whether the original evidence was adequately removed, and whether the background edit avoided clinically relevant regions. Images with obvious editing artifacts were regenerated. The final visual-obviousness rejection rate was 7.3% for radiographs and 10.8% for dermatology images. Dermatology edits were more difficult because skin texture, lighting, and lesion boundaries can make inpainting visible. The final edited images were not intended to deceive specialists under forensic inspection; they were intended to remove or preserve evidence without creating gross artifacts that would dominate model behavior.

The relocation variant required careful definition. In radiographs, a fracture-like region could be moved to a nonbone region or to a different anatomical site where the original answer would no longer be valid [7]. A pneumothorax boundary could be relocated away from the pleural margin, creating a pattern that should not support the same answer. In dermatology images, a lesion subregion could be moved to surrounding normal skin or duplicated in a way that disrupts the original border explanation. The relocation variant was included because erasure alone cannot distinguish a model that notices evidence absence from one that simply reacts to any edit. Relocation tests whether the model tracks the region's meaning in context.

The editing process produced four images per case: original, lesion-erased, background-erased, and relocated. With 5,760 cases, the image set contained 23,040 images. Each image was paired with the original question. Five multimodal systems were evaluated, producing 115,200 answer-explanation outputs before exclusions. A small number of outputs were removed because of corrupted image loading or blank model response [8]. The final analysis included 114,486 outputs. Missing outputs were distributed across models and conditions without meaningful concentration in one modality.

The prompt asked the model to provide a direct answer followed by one short reason based on visible image evidence. The wording did not request long reasoning or chain-of-thought. The answer format was intentionally compact because the evaluation concerns whether the explanation changes appropriately, not whether the model can produce a full report. The same prompt was used for all image variants. No prompt stated that the image had been altered. The models therefore had no textual reason to discuss editing.

The output parser separated the direct answer from the explanatory reason. For example, if a response said yes, there is a small right apical pneumothorax because a pleural line is visible with absent peripheral markings, the direct answer was yes, and the explanation contained pleural line and absent peripheral markings. The parser used punctuation, discourse markers, and a radiology-dermatology phrase dictionary. A sample of 2,000 parsed outputs was manually checked. Parser accuracy for separating answer and explanation was 93.1%. Parser mistakes were corrected in the human-audited subset and included as ordinary system errors elsewhere.

Expert labels of explanation groundedness were collected for a stratified sample of 8,000 original-image outputs.

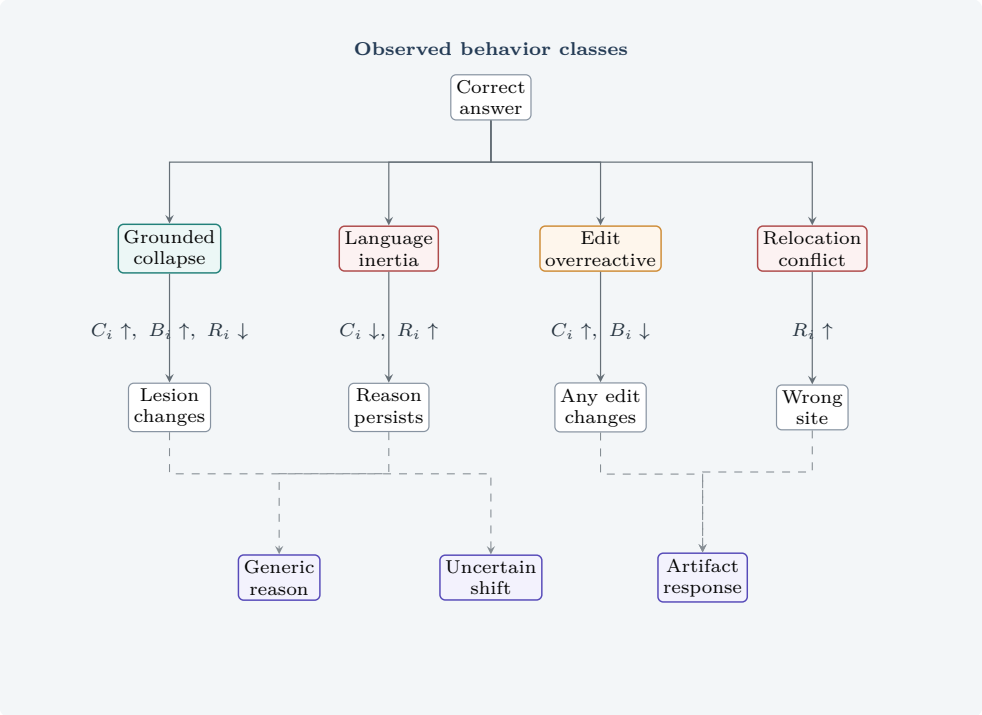

Figure 5: Failure-mode taxonomy for correct original answers. Counterfactual edits separate explanations that truly respond to lesion evidence from outputs dominated by language templates, general edit sensitivity, spatial inconsistency, or unsupported generic phrasing.

Reviewers judged whether the explanation referred to an image feature that was actually present in the masked evidence region, whether it was too generic, whether it mentioned an incorrect region, and whether it relied on nonvisual clinical assumptions. The main binary label was grounded versus not grounded. A grounded explanation had to name or imply a visible feature in the correct region. Merely saying the image shows it was not sufficient. Inter-reviewer agreement was 0.82 for the binary groundedness label.

The edited-condition outputs were graded for expected counterfactual behavior. For lesion-erased variants, reviewers judged whether abnormality-specific explanation collapsed, weakened, or inappropriately persisted. For background-erased variants, reviewers judged whether the explanation remained stable or changed without reason. For relocated variants, reviewers judged whether the model recognized the changed visual context or contradicted the edited image. These labels were used to validate automatic metrics and to train the latent grounding coefficient on the training partition.

The dataset was divided into 60% training, 20% validation, and 20% test sets by patient or case group. All headline results use the held-out test set. The training partition was used only for fitting automatic scoring components and the latent grounding model. The base vision-language models were not fine-tuned. The validation partition was used for threshold selection, model hyperparameters, and grader normalization. No image from the test set was used in controller fitting, threshold selection, or prompt revision.

3 Counterfactual Grounding Model

The mathematical goal is to estimate whether an explanation depends on the masked visual evidence. Let x_i be the original image, x_i^- the lesion-erased image, x_i^b the background-erased image, and x_i^r the relocated image. Let q_i be the question and let $y_i^0, y_i^-, y_i^b, y_i^r$ be the corresponding model outputs. Each output is decomposed into an answer component a and an explanation component e. The model seeks a grounding coefficient $g_i \in [0, 1]$, where higher values indicate stronger visual dependence of the original explanation on the lesion region.

The first metric is counterfactual explanation collapse. The intuition is that explanation content specific to the original lesion should decrease after lesion erasure. Let $\phi(e)$ be an embedding of the explanation, and let $\psi(e)$ be a vector of abnormality-specific visual terms. Let $s(e)$ be a scalar abnormality support score estimated by an explanation classifier. Collapse is defined as

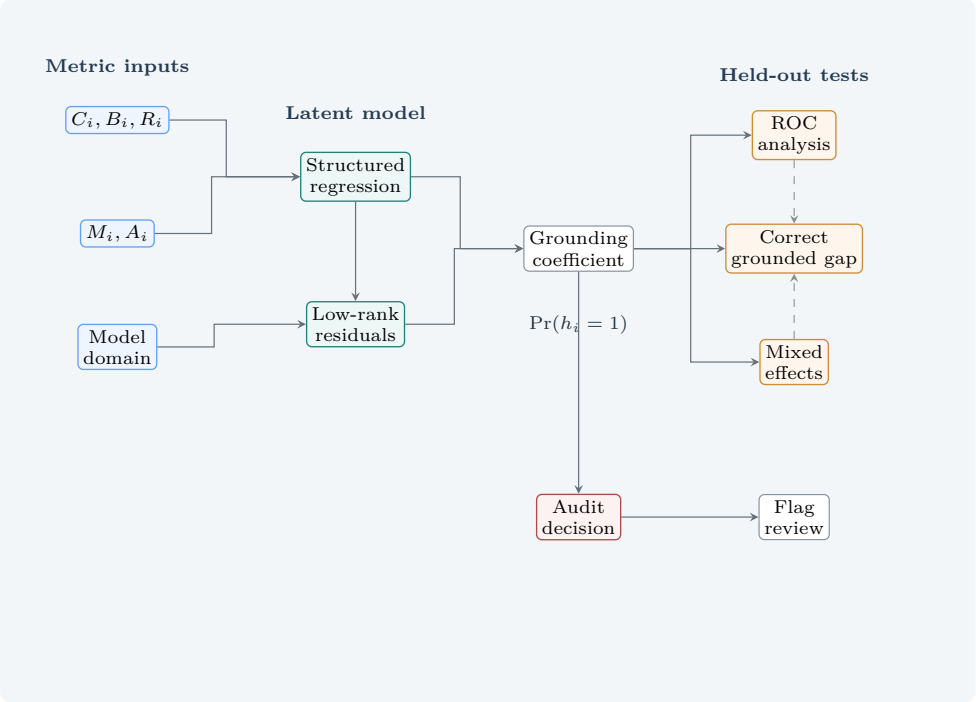

Figure 6: Statistical analysis graph. Counterfactual metrics, alignment, correctness, and model-domain indicators are fused through a structured latent model, then evaluated against expert groundedness labels and used to estimate the gap between ordinary accuracy and grounded-correct performance.

$$C_i = \left[s(e_i^0) - s(e_i^-)\right]_+ \cdot \left(1 - \frac{\phi(e_i^0)^\top \phi(e_i^-)}{\|\phi(e_i^0)\|_2 \|\phi(e_i^-)\|_2}\right)^{1/2}. \tag{1}$$

The positive part ensures that collapse is credited only when abnormality support decreases after lesion erasure. The embedding term prevents a superficial drop in one keyword from being treated as full collapse if the explanation remains semantically the same [9]. A high value of C_i indicates that the explanation changed in the expected direction after evidence removal.

The second metric is region-preservation stability. A grounded explanation should not change much when an irrelevant matched region is erased. Stability is defined as

$$B_i = \frac{\phi(e_i^0)^\top \phi(e_i^b)}{\|\phi(e_i^0)\|_2 \|\phi(e_i^b)\|_2} \cdot \left(1 - |s(e_i^0) - s(e_i^b)|\right). \tag{2}$$

The first factor measures semantic similarity between original and background-erased explanations. The second factor penalizes changes in abnormality support. High B_i indicates that irrelevant edits did not disturb the explanation. This metric helps distinguish true evidence sensitivity from general edit sensitivity. A model that changes its answer after any inpainting would have low B_i.

The third metric is relocation contradiction [10]. If the lesion evidence is moved to a context where the original explanation is no longer anatomically or clinically valid, the model should not repeat the same explanation as though nothing changed. Let R_i be a contradiction score between the relocated explanation and the edited image context. Since direct image-text contradiction is difficult, R_i combines an anatomical consistency classifier with explanation similarity to the original output:

$$R_i = \mathrm{sim}(e_i^0, e_i^r) \cdot (1 - \mathrm{cons}(e_i^r, x_i^r, q_i)). \tag{3}$$

Here, sim is embedding cosine similarity, and cons estimates whether the relocated explanation is anatomically consistent with the relocated image. High R_i means the model preserved the original explanation despite inconsistent relocation, which is undesirable. For the final grounding score, relocation contradiction is subtracted [11].

The three metrics are combined with answer correctness under the original image. The model should not give high grounding credit to an explanation attached to a wrong answer, even if the text changes under erasure. Let A_i be the strict correctness indicator for the original answer. The raw counterfactual grounding score is

$$G_i^{raw} = A_i \cdot \sigma\left(\theta_0 + \theta_C C_i + \theta_B B_i - \theta_R R_i + \theta_M M_i\right). \tag{4}$$

The term M_i measures mask-language alignment, described below. Multiplication by A_i forces wrong original answers toward low grounding in the primary version. A secondary analysis removes this multiplication to examine grounding independent of correctness.

Mask-language alignment estimates whether the original explanation refers to the masked region. The model uses a phrase-to-region alignment matrix. Suppose an explanation contains K_i visual phrases and the image is divided into P_i regions. Let $Z_i \in \mathbb{R}^{K_i \times P_i}$ contain phrase-region compatibility scores from a cross-modal grounding model. Let $m_i \in \{0,1\}^{P_i}$ indicate whether each region overlaps the expert mask. The alignment score is

$$M_i = \frac{1}{K_i} \sum_{k=1}^{K_i} \frac{\sum_{p=1}^{P_i} Z_{ikp} m_{ip}}{\sum_{p=1}^{P_i} Z_{ikp} + \epsilon}. \tag{5}$$

A high M_i means the explanation's visual phrases align with the expert-marked region. Attention maps alone can produce this type of score, but the present model does not rely on it alone because a model may attend to the right region without making its text depend on that region. The counterfactual metrics are therefore primary, and mask-language alignment is auxiliary.

The latent grounding coefficient is estimated through a structured regression model trained on expert groundedness labels. Let h_i be the binary expert label for whether the original explanation is grounded. The coefficient is

$$\Pr(h_i = 1) = \sigma\left(\beta_0 + \beta_1 C_i + \beta_2 B_i - \beta_3 R_i + \beta_4 M_i + \beta_5 A_i + u_{m(i)} + v_{d(i)}\right). \tag{6}$$

Here, $u_{m(i)}$ is a model-specific effect and $v_{d(i)}$ is a domain-specific effect for radiography or dermatology. This formulation allows some models to have higher baseline groundedness and allows domains to differ in visual-text behavior. The parameters are fitted using the training partition, tuned on validation, and evaluated on the held-out test partition.

A matrix version of the model is used to analyze patterns across findings. Let $Y \in \mathbb{R}^{N \times 3}$ contain the three counterfactual metrics C, B, R. Let $X \in \mathbb{R}^{N \times L}$ contain finding-type indicators and modality indicators. Let $W \in \mathbb{R}^{L \times 3}$ map finding types to expected counterfactual behavior. The model estimates

$$\begin{aligned} Y &= XW + \Gamma + E, \\ \Gamma &= PQ^\top, \quad P \in \mathbb{R}^{N \times r}, \quad Q \in \mathbb{R}^{3 \times r}. \end{aligned} \tag{7}$$

The low-rank term Γ captures residual structure not explained by finding type, such as model-specific tendency to repeat explanations. In the final analysis, rank $r = 2$ was sufficient. The first residual factor corresponded to language persistence across edits. The second corresponded to general sensitivity to inpainting artifacts.

For statistical testing, the main estimand is the difference between correct-answer rate and grounded-correct rate. Correct-answer rate measures how often the original response answer is correct. Grounded-correct rate measures how often the original answer is correct and the explanation receives a grounding coefficient above a selected threshold. Let A_i be correctness and H_i be groundedness. The difference is

$$D = \frac{1}{n} \sum_{i=1}^{n} A_i - \frac{1}{n} \sum_{i=1}^{n} A_i H_i. \tag{8}$$

A large D indicates that many correct answers are not counterfactually grounded. Confidence intervals are computed by case-level bootstrap. The grounding threshold is selected on validation to maximize balanced accuracy against expert labels.

The model also estimates a collapse asymmetry ratio. For a grounded explanation, erasing the lesion should produce larger change than erasing background. The ratio is

$$\Lambda_i = \frac{1 - \mathrm{sim}(e_i^0, e_i^-)}{1 - \mathrm{sim}(e_i^0, e_i^b) + \epsilon}. \tag{9}$$

Values above one indicate lesion-specific change. Values near one indicate that the model changed similarly for lesion and background edits. Values below one suggest that background erasure disturbed the explanation more than lesion erasure. The ratio is not stable when both similarities are near one, so it is used as a secondary descriptive statistic rather than a primary classifier.

Finally, answer-explanation coupling is quantified. Some models change the explanation but keep the answer, while others change both. Let Δa_i be answer change under lesion erasure and Δe_i be explanation change. Coupling is

$$K_i = \operatorname{corr}(\Delta a_i, \Delta e_i) \quad \text{within} \quad \mathcal{G}_{m,d}, \tag{10}$$

where $\mathcal{G}_{m,d}$ groups outputs by model and domain. High coupling means the answer and explanation respond together. Low coupling means the explanation may change decoratively while the answer remains fixed, or the answer changes without a coherent explanation shift. Coupling is reported by model and modality.

4 Experimental Procedure

Five multimodal systems were evaluated. They are named Model One through Model Five. Two systems were broad multimodal instruction models, two were medical-domain models, and one was a compact locally hosted model. The models were not trained or adapted on the benchmark. Every model received the same image and the same question for each of the four image conditions. Decoding used deterministic settings where available. The maximum output was 80 tokens because the task asked for a short answer and one reason. Outputs longer than 80 tokens were truncated only for storage; the full raw output was retained for grading when available.

The evaluation used four image conditions per case. The original condition established baseline answer correctness and explanation content. The lesion-erased condition measured collapse. The background-erased condition measured stability under irrelevant edit. The relocated condition measured context-sensitive explanation change. The order of model queries was randomized so that conditions from the same case did not appear consecutively for systems with possible conversational memory. Each query was submitted as an independent session where the interface allowed it. This avoided carryover from one image condition to another.

The output grading process had automatic and human components. Answer correctness was determined by a radiology-dermatology semantic grader trained on normalized reference answers and audited by domain reviewers. Explanation collapse and stability were scored by automatic comparison metrics and then validated against human labels. Human reviewers did not know which model produced a response. For edited images, reviewers saw the edited image and the response but did not see the original response unless the task required judging persistence. For persistence-like judgments in relocation, reviewers saw both original and relocated outputs with model identity hidden.

The primary human audit included 12,000 outputs, with balanced sampling across models, modalities, and image conditions. The audit oversampled cases where automatic metrics and answer correctness disagreed. For example, cases with a correct answer but low collapse were oversampled because they are central to the paper's claim. Cases with high collapse but wrong original answer were also sampled to understand whether the collapse metric could be fooled by unstable language. Inter-reviewer agreement was 0.84 for answer correctness, 0.78 for explanation collapse category, 0.81 for background stability, and 0.75 for relocation contradiction. The relocation task was harder because reviewers had to judge anatomical consistency after edited evidence movement.

The expert groundedness label for original explanations was used as the main external standard for the latent grounding coefficient. This label asked whether the explanation pointed to evidence actually visible in the original image and localized within the expert mask. A correct but generic explanation such as because the abnormality is visible did not receive grounded credit. A statement naming a finding without a visual reason also did not receive grounded credit. A statement mentioning a visual sign in the wrong region was marked ungrounded. The standard was intentionally strict because the paper studies explanation evidence, not answer acceptability.

The primary quantitative comparison evaluated six scoring approaches. The first was answer correctness alone. The second was explanation attention overlap using phrase-region alignment without counterfactual edits. The third was semantic similarity between original and lesion-erased explanations. The fourth was collapse score C_i alone. The fifth was the combination of collapse and background stability. The sixth was the full latent grounding coefficient including collapse, stability, relocation contradiction, mask-language alignment, correctness, and model-domain effects. These approaches were compared against expert groundedness labels using area under the receiver operating characteristic curve, area under the precision-recall curve, calibration error, and balanced accuracy.

Statistical analysis was performed at the case level. Since each case produced outputs from multiple models and conditions, standard errors were clustered by case. Model comparisons used mixed-effects logistic regression with random intercepts for case and fixed effects for model, modality, finding type, and image condition. The main regression for explanation persistence after lesion erasure was

$$\operatorname{logit} \Pr(P_{ij} = 1) = \alpha + \mu_{m(j)} + \nu_{d(i)} + \omega_{f(i)} + \zeta A_{ij} + b_i, \tag{11}$$

where P_{ij} indicates inappropriate persistence of abnormality explanation for case i and model j, $m(j)$ is model identity, $d(i)$ is domain, $f(i)$ is finding type, A_{ij} is original answer correctness, and b_i is a random case effect. This model estimates whether persistence differs across models after accounting for case and finding factors.

The study also examined whether counterfactual behavior could be predicted from original output only. A baseline classifier used original answer text, original explanation text, model identity, and question type but did not use edited image outputs. This baseline represents a cheaper evaluation that avoids counterfactual generation. Its performance was compared with the full counterfactual coefficient. The comparison tests whether direct intervention on the image adds information beyond textual style.

Thresholds for groundedness were selected on validation data. The threshold for the full grounding coefficient was chosen to maximize the mean of sensitivity and specificity against expert groundedness. A secondary threshold was chosen to favor sensitivity, suitable for audit settings where ungrounded explanations should be flagged for review. All test results report both the balanced threshold and the high-sensitivity threshold. The headline grounded-correct rate uses the balanced threshold.

To examine modality differences, radiographic and dermatologic outputs were analyzed separately. Radiographic abnormalities often occupy smaller regions, and erasure may create subtle changes. Dermatologic lesions occupy larger portions of the image, and erasure may alter the entire visual focus. The study therefore expected higher collapse in dermatology but also higher sensitivity to inpainting artifacts. The background-erasure condition was important for testing this expectation because dermatology background edits may affect lighting or skin texture in ways that disturb the model.

To examine finding differences, cases were grouped into seven finding families: pleural abnormality, parenchymal opacity, fracture or dislocation, device or hardware, pigmentation pattern, inflammatory surface change, and ulceration or focal lesion defect. Each family had different evidence structure. Device and hardware findings often have sharp edges and strong visual cues. Pigmentation and inflammatory findings have texture and color gradients. Pleural abnormalities can be subtle and anatomy-dependent. The counterfactual metrics were reported by family.

The study included a negative-control analysis. In negative cases with no target abnormality, erasing the assigned salient but irrelevant region should not cause a model to newly assert the abnormality. If erasure produces new abnormality language, the model is responding to edit artifacts or unstable priors. Negative-control instability was calculated as the rate at which an originally correct negative answer became positive after lesion-region erasure or background erasure. This analysis helped identify models that were generally unstable under image edits.

A second control analysis used mask dilation and contraction. For a subset of 900 cases, the expert mask was expanded by 25% and contracted by 25%. Erasure was repeated with both modified masks. A truly localized explanation should show stronger collapse with the original or dilated mask than with the contracted mask when the contracted mask misses relevant evidence. This sensitivity analysis evaluated whether the results depended on exact mask boundaries. It also helped identify cases where the expert mask was too narrow.

5 Results

Table 1: Dataset Composition and Annotation Statistics

Category	Cases	Median Mask Area	Notes
Radiographic images	3,360	4.8%	Chest and extremity findings
Dermatology images	2,400	18.6%	Clinical lesion photographs
Positive abnormality cases	Oversampled	–	Used for collapse analysis
Negative-control cases	Included	Salient regions	Artifact sensitivity testing
Edited image variants	23,040	–	Four variants per case
Final analyzed outputs	114,486	–	After corrupted-output removal

Table 2: Counterfactual Image Conditions

Condition	Transformation	Intended Effect	Evaluation Focus
Original	None	Baseline response	Accuracy and grounding
Lesion-erased	Mask inpainting	Remove evidence	Explanation collapse
Background-erased	Irrelevant region erased	Preserve evidence	Stability assessment
Relocated evidence	Evidence repositioned	Break anatomy consistency	Relocation contradiction

Table 3: Primary Counterfactual Metrics

Metric	Desired Trend	Interpretation	Main Failure Type
Collapse score (C_i)	High	Explanation weakens after erasure	Language persistence
Background stability (B_i)	High	Stable under irrelevant edits	Edit sensitivity
Relocation contradiction (R_i)	Low	Avoids inconsistent reuse	Spatial inconsistency
Mask alignment (M_i)	High	Phrase overlaps evidence region	Weak localization
Grounding coefficient	High	Combined evidence dependence	Ungrounded reasoning

Table 4: Overall Model Performance

Model	Accuracy	Grounded-Correct Rate	Coupling K
Model One	83.4%	50.8%	0.62
Model Two	80.2%	47.3%	0.57
Model Three	78.5%	42.8%	0.43
Model Four	76.8%	38.9%	0.36
Model Five	74.1%	34.6%	0.31

Original-image strict answer accuracy across all models and cases was 78.6%. Radiographic accuracy was 76.9%, and dermatologic accuracy was 81.0%. Model One had the highest original accuracy at 83.4%, followed by Model Two at 80.2%, Model Three at 78.5%, Model Four at 76.8%, and Model Five at 74.1%. Accuracy varied by finding family. Device and hardware questions had the highest accuracy at 84.6%, while pleural abnormalities had the lowest at 72.3%. Dermatologic pigmentation-pattern questions reached 82.4%, while ulceration-focused questions reached 79.1%.

Correct answer accuracy overstated grounded explanation behavior. Among correct original-image answers, only 54.2% showed expected explanation collapse after lesion erasure. In 29.7% of correct answers, the explanation persisted with little change even after the expert-marked evidence was erased. In 16.1%, the response changed but not in a clinically coherent direction. Some of these cases replaced one unsupported explanation with another, such as shifting from visible pleural line to apical lucency without acknowledging that the pleural evidence was removed. This gap between correctness and collapse is the main empirical finding.

Background-erasure stability was higher than lesion-erasure collapse but still imperfect. Among correct original answers, 71.5% preserved explanation content when irrelevant background was erased. The remaining 28.5% changed in ways that suggested general image-edit sensitivity. Dermatology images had lower background stability than radiographs, at 66.2% compared with 75.1%. Reviewers attributed many dermatology instability cases to lighting and texture changes introduced by background inpainting. This result shows why lesion-erasure collapse must be interpreted alongside background stability. A model that changes after lesion erasure but also changes after irrelevant erasure may not be truly evidence-sensitive.

Relocation contradiction occurred in 32.8% of originally correct explanations. In these cases, the model repeated or preserved the original explanation after the evidence was relocated to an inconsistent region. The rate was highest for dermatologic border irregularity at 39.5% and lowest for orthopedic hardware at 21.7%. Radiographic fracture relocation produced contradictions in 28.4%. Relocation failures often looked fluent. The model might continue to describe a distal radial cortical break after the fracture-like patch was moved away from the radius, or it might continue to describe asymmetry of the original skin lesion after the irregular border segment had been relocated. These outputs suggest that the model's explanation was attached to the original question-answer pattern more than the edited image.

The full latent grounding coefficient achieved area under the receiver operating characteristic curve of 0.817 against expert groundedness labels. Area under the precision-recall curve was 0.744. Answer correctness alone achieved area under the receiver operating characteristic curve of 0.641. Attention-style mask-language alignment alone achieved 0.692. Original-to-erased semantic similarity alone achieved 0.713. Collapse score alone achieved 0.742. Collapse plus background stability achieved 0.781. The full coefficient performed best, indicating that no single signal captured the whole phenomenon. Relocation contradiction and background stability added information beyond collapse.

At the balanced validation threshold, the full coefficient reached sensitivity of 0.766 and specificity of 0.741 on the held-out test set. At the high-sensitivity threshold, sensitivity increased to 0.884 while specificity decreased to 0.592. The high-sensitivity threshold may be appropriate for auditing generated explanations before release. The balanced threshold was used for estimating grounded-correct rates. Under this threshold, the overall grounded-correct rate was 43.1%, compared with ordinary correctness of 78.6%. This means that a large share of correct answers either lacked sufficiently grounded explanations or failed counterfactual evidence tests.

Model differences were substantial. Model One had the highest grounded-correct rate at 50.8%, while Model Five had the lowest at 34.6%. The ranking was similar but not identical to ordinary accuracy. Model Two

Table 5: Grounding Outcomes Across Conditions

Outcome	Overall Rate	Radiography	Dermatology
Correct answer accuracy	78.6%	76.9%	81.0%
Expected collapse	54.2%	49.8%	59.7%
Background stability	71.5%	75.1%	66.2%
Relocation contradiction	32.8%	28.4%	39.5%
Grounded-correct outputs	43.1%	40.4%	46.5%

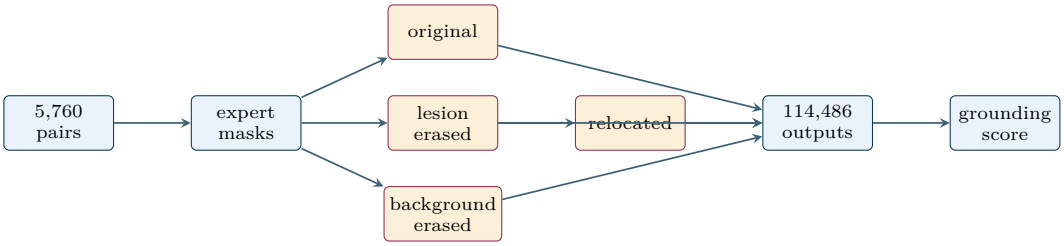

Figure 7: Counterfactual evaluation flow from masked image-question pairs to edited variants, model outputs, and the final latent grounding score.

had lower accuracy than Model One but similar collapse behavior among correct answers. Model Three had moderate accuracy but poor relocation behavior. Model Four changed explanations frequently under both lesion and background erasure, suggesting edit sensitivity rather than evidence specificity. Model Five often produced generic explanations that were hard to ground, such as because the image findings support it.

Radiography and dermatology showed different counterfactual profiles. Dermatology had higher original accuracy and stronger lesion-erasure collapse, with mean collapse score 0.412 compared with 0.337 for radiography. However, dermatology also had lower background stability and higher relocation contradiction for border and color explanations. Radiography had weaker collapse but better background stability. These patterns suggest that dermatologic models react strongly to visual changes but may not always distinguish lesion evidence from general image texture. Radiographic models are more stable but may preserve learned explanation phrases even when subtle evidence is erased [12].

Finding-family analysis revealed that device and hardware explanations were most grounded. Among correct device or hardware answers, 68.9% passed the groundedness threshold. Fracture and dislocation explanations passed at 51.7%. Pleural abnormality explanations passed at 38.6%. Parenchymal opacity explanations passed at 42.9%. Dermatologic pigmentation explanations passed at 48.2%, inflammatory surface-change explanations at 45.6%, and ulceration or focal defect explanations at 52.1%. Pleural findings were particularly difficult because models often used stock phrases about apical lucency or pleural line even after the marked evidence was removed.

The negative-control analysis showed that edits could induce false positives. In cases with no target abnormality, lesion-region erasure caused a previously correct negative answer to become positive in 6.8% of outputs. Background erasure caused such conversion in 5.9%. The difference was small, indicating that some models respond to editing artifacts generally. Model Four had the highest negative-control instability at 9.7%, while Model One had the lowest at 3.8%. Dermatology negative controls were more unstable than radiographic negative controls. This supports the need to inspect editing artifacts and to include background-erasure controls.

Mask dilation and contraction produced expected but not perfect trends. Dilated-mask erasure increased mean collapse by 7.4% relative to original-mask erasure. Contracted-mask erasure reduced mean collapse by 11.6%. The effect was strongest for dermatologic lesions, where masks covered visible lesion boundaries. Radiographic effects were smaller because some findings had diffuse evidence outside the mask. In pleural effusion and parenchymal opacity cases, contracted masks sometimes still removed enough evidence to affect explanation. These results show that mask quality matters but does not fully explain the observed grounding gap.

The collapse asymmetry ratio Λ_i was informative. Grounded expert-labeled explanations had median $\Lambda_i = 2.41$, meaning lesion erasure caused more than twice the explanation change caused by background erasure. Ungrounded explanations had median $\Lambda_i = 1.08$, suggesting nearly equal response to lesion and background edits. The ratio was especially discriminative for fracture and pigmentation cases. It was less discriminative for diffuse opacity cases, where background regions sometimes contained contextual lung texture that affected model output.

Answer-explanation coupling differed by model. Model One had the highest coupling between answer change and explanation change under lesion erasure, with $K = 0.62$. Model Two had $K = 0.57$, Model Three had $K = 0.43$, Model Four had $K = 0.36$, and Model Five had $K = 0.31$. Low coupling often meant that the model kept the same answer while changing explanation wording, or changed the answer without providing an explanation that reflected the visual alteration. Higher coupling was associated with higher grounded-correct rate, but the relationship was

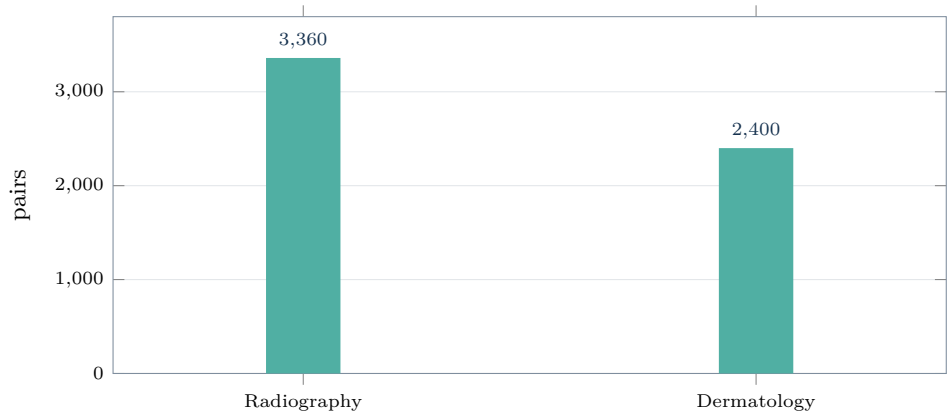

Figure 8: Benchmark composition includes 3,360 radiographic and 2,400 dermatologic image-question pairs, giving 5,760 total masked cases.

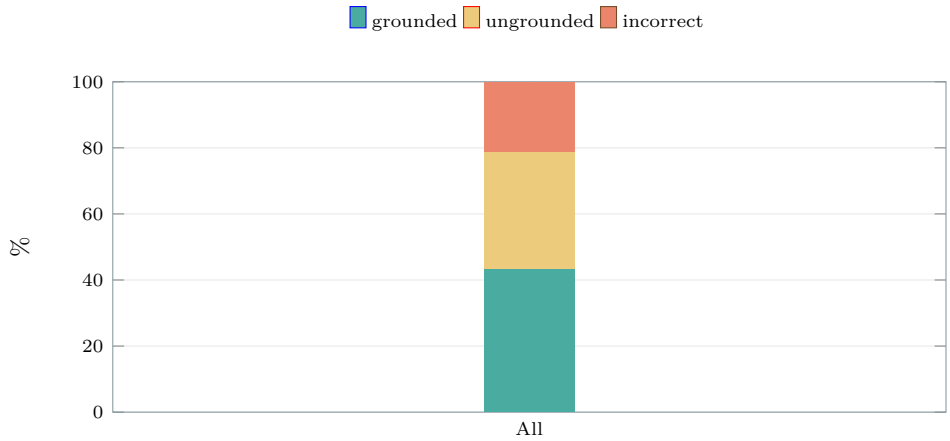

Figure 9: Ordinary accuracy was 78.6%, but only 43.1% of all outputs were both correct and counterfactually grounded under the balanced threshold.

not perfect. Some models changed both answer and explanation for irrelevant edits, which increased coupling without improving specificity.

The original-output-only baseline performed worse than the full counterfactual coefficient. It achieved area under the receiver operating characteristic curve of 0.701. It could identify obviously generic explanations but failed on fluent ungrounded explanations that named plausible signs. For example, if a model said visible cortical discontinuity supports fracture, the text alone looked grounded. Only after erasing the fracture region could the evaluator see whether the explanation depended on that evidence. This confirms that counterfactual image intervention adds information not present in original output style.

Mixed-effects regression showed that inappropriate persistence after lesion erasure was strongly associated with model identity and finding type. After controlling for original answer correctness, pleural abnormalities had 1.74 times higher odds of persistence than device or hardware findings. Dermatologic border irregularity had 1.49 times higher odds than ulceration. Model Five had 2.08 times higher odds of persistence than Model One. All three differences remained significant after false discovery rate correction. Original answer correctness itself increased persistence odds slightly because correct original answers often came with more confident explanation templates.

The full grounding coefficient was reasonably calibrated. Expected calibration error was 0.064 overall, 0.059 for radiography, and 0.072 for dermatology. Calibration was weaker for dermatologic inflammatory surface-change cases, where explanations often referred to redness, scale, and texture in overlapping ways. The coefficient tended to overestimate groundedness when explanations contained multiple visual terms but only one term was actually mask-supported. Predicate-level scoring may address this issue in future work.

Human reviewers judged that 18.6% of explanations were generic even when the answer was correct. Generic explanations were especially common in Model Five and in radiographic opacity cases. Generic explanation did not always mean unsafe output, but it reduced interpretability. When a user asks for a reason, because the image shows it does not provide meaningful evidence. Counterfactual erasure rarely produced clean collapse in these cases because there was little specific content to collapse. The grounding coefficient correctly assigned low values to most generic explanations.

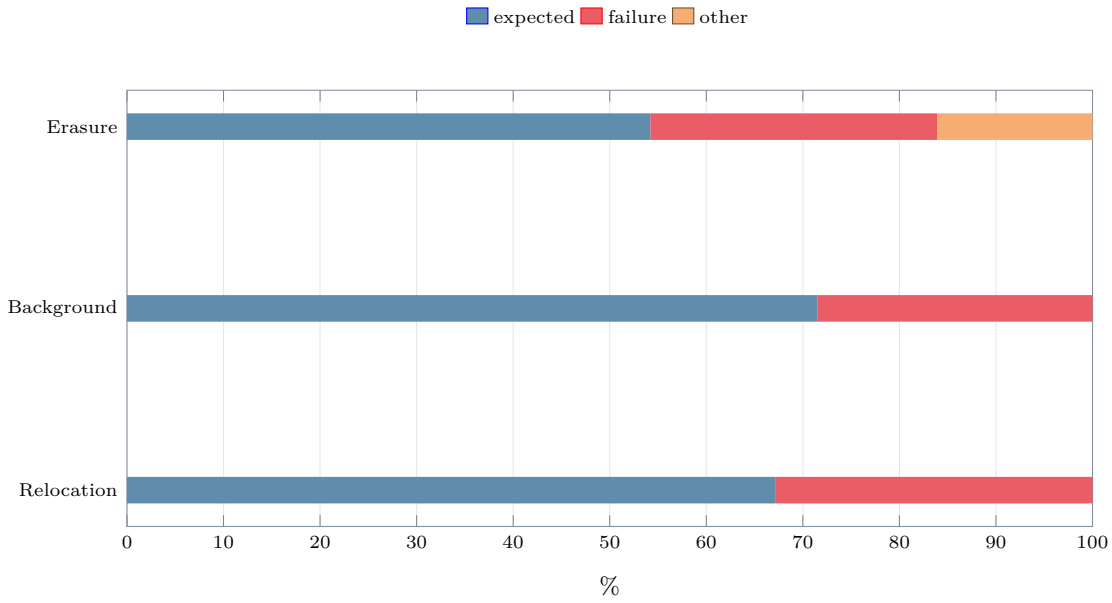

Figure 10: Counterfactual checks separate expected lesion collapse, background stability, and relocation consistency from persistence, edit sensitivity, and incoherent shifts.

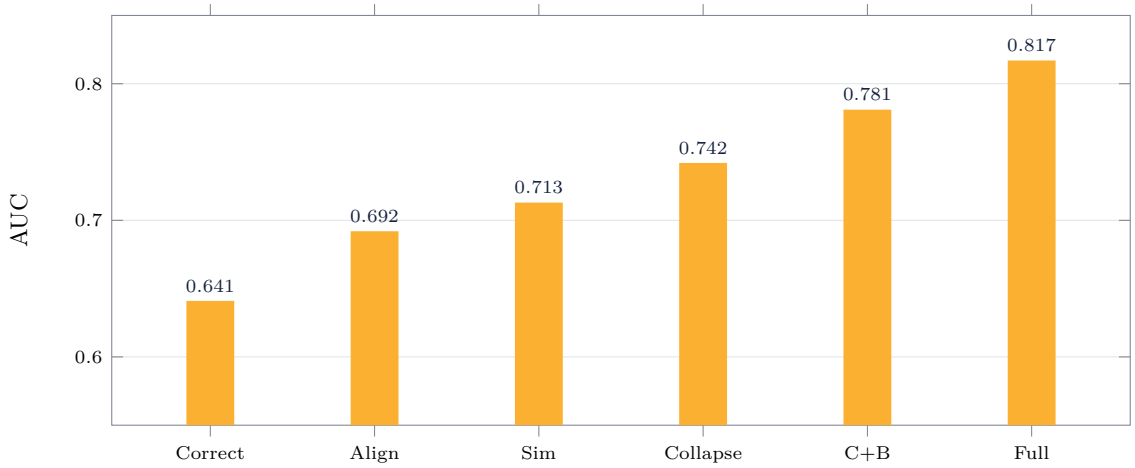

Figure 11: The full latent coefficient outperformed correctness, static alignment, semantic similarity, collapse alone, and collapse plus background stability.

The relocation condition revealed a failure that erasure alone missed. In 9.3% of cases, the explanation collapsed appropriately after lesion erasure but still contradicted the relocated image. These cases would look grounded under erasure-only evaluation. The model noticed when evidence was absent but did not properly reinterpret evidence when moved. This was common in dermatology, where the model sometimes continued to discuss the original lesion boundary after the boundary segment was relocated. It also occurred in fractures when the model recognized that a fracture-like feature existed but described the wrong anatomical site. Relocation therefore adds a spatial-consistency test that erasure cannot provide.

6 Extended Mathematical Analysis

The counterfactual framework can be understood through a causal response model. Let V_i represent the visual evidence in the expert mask and L_i represent language priors derived from question and model training. The explanation E_i is generated from both sources. A simplified structural form is

$$E_i = F_\theta \left(V_i, L_i, U_i \right), \tag{12}$$

where U_i represents unobserved randomness or hidden model state. A visually grounded explanation should

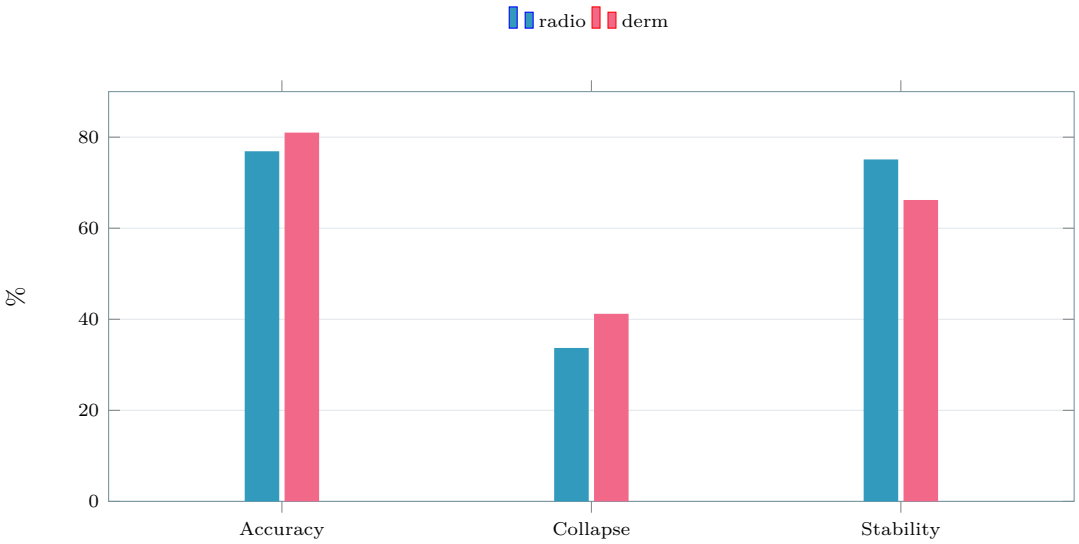

Figure 12: Dermatology showed higher original accuracy and stronger lesion-erasure collapse, while radiography showed higher background-erasure stability.

Table 6: Finding Family Groundedness Results

Finding Family	Accuracy	Grounded Rate	Key Observation
Device and hardware	84.6%	68.9%	Strong visual cues
Fracture/dislocation	78.2%	51.7%	Moderate collapse behavior
Pleural abnormality	72.3%	38.6%	High template persistence
Parenchymal opacity	75.4%	42.9%	Diffuse evidence challenge
Pigmentation pattern	82.4%	48.2%	Texture-sensitive responses
Inflammatory surface change	79.8%	45.6%	Background instability present

have a nonzero causal dependence on V_i. Lesion erasure approximates an intervention $do(V_i = 0)$. The observed explanation difference estimates the effect of removing visual evidence while holding the question fixed. Although the model internals are not accessible, the output response to intervention can reveal dependence.

The average counterfactual effect of lesion evidence on explanation support is

$$\mathrm{ACE}_E = \mathbb{E}\left[s(E_i) \mid do(V_i = V_i^{obs})\right] - \mathbb{E}\left[s(E_i) \mid do(V_i = 0)\right]. \tag{13}$$

In practice, $s(E_i)$ is estimated from text, and the do-intervention is approximated by inpainting. This is not a perfect intervention because erasure may change other image statistics. The background-erasure condition estimates the nuisance effect of image editing. A corrected causal effect can be written as

$$\begin{aligned} \mathrm{ACE}_{corr} &= \left[s(e_i^0) - s(e_i^-)\right] - \left[s(e_i^0) - s(e_i^b)\right] \\ &= s(e_i^b) - s(e_i^-). \end{aligned} \tag{14}$$

This corrected effect is large when lesion erasure reduces abnormality support more than background erasure. It is small when both edits have similar effects. The collapse metric C_i can be viewed as a semantic version of this corrected effect combined with explanation distance.

The framework can also be interpreted using projection onto evidence subspaces. Let r_i be a representation of the image region inside the mask and o_i a representation of regions outside the mask. Let t_i be a representation of the explanation. A grounded explanation should have higher projection onto the masked evidence than onto irrelevant background. Define

$$\begin{aligned} P_{r_i} &= R_i \left(R_i^\top R_i + \lambda I\right)^{-1} R_i^\top, \\ P_{o_i} &= O_i \left(O_i^\top O_i + \lambda I\right)^{-1} O_i^\top, \\ \Omega_i &= \|P_{r_i} t_i\|_2^2 - \|P_{o_i} t_i\|_2^2. \end{aligned} \tag{15}$$

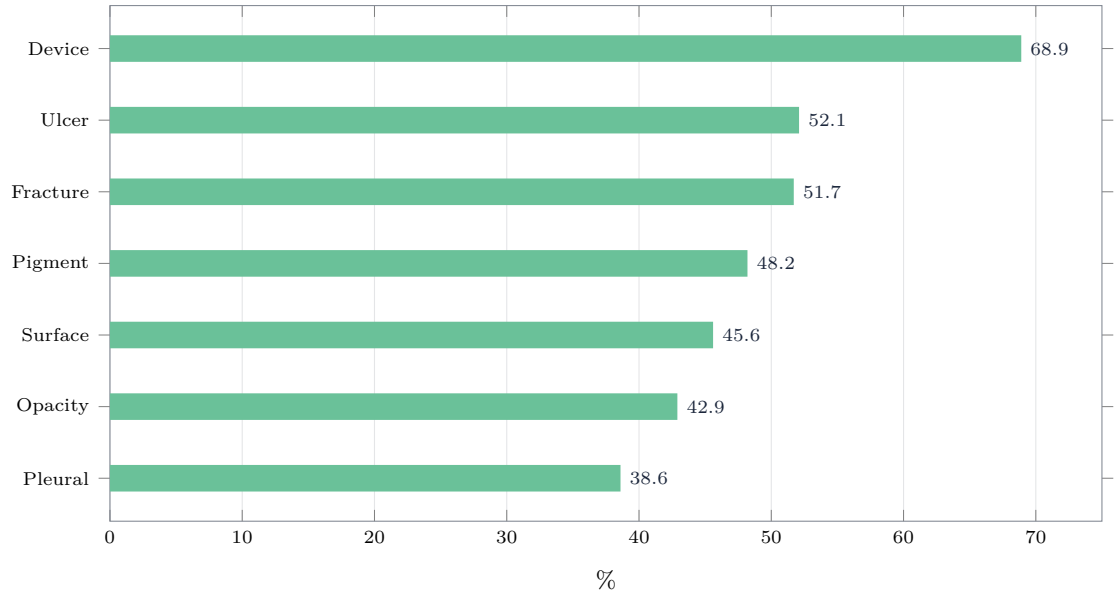

Figure 13: Grounded-correct rates varied by finding family, with device or hardware explanations strongest and pleural abnormality explanations weakest.

Table 7: Evaluation Method Comparison

Method	AUROC	AUPRC	Relative Strength
Answer correctness	0.641	–	Weak grounding proxy
Mask-language alignment	0.692	–	Static localization only
Semantic similarity	0.713	–	Limited intervention signal
Collapse score only	0.742	–	Sensitive to evidence removal
Collapse + stability	0.781	–	Better edit discrimination
Full grounding coefficient	0.817	0.744	Best overall performance

A positive Ω_i suggests that explanation content aligns more with masked evidence than with background. However, this static alignment is weaker than counterfactual testing because it does not show whether the explanation changes when evidence changes. In the experiments, projection-style alignment improved over correctness alone but underperformed the full coefficient.

A more detailed model treats explanation phrases as predicates. Suppose explanation e_i contains predicates $p_{i1}, \ldots, p_{iK}$, such as pleural line visible, border irregular, cortical break present, or color variation. Each predicate has a visual support probability z_{ik}. After lesion erasure, support should decline for predicates tied to the mask. Predicate-level collapse is

$$C_i^{pred} = \frac{1}{K} \sum_{k=1}^{K} w_{ik} \left[z_{ik}^{0} - z_{ik}^{-} \right]_+ , \tag{16}$$

where w_{ik} weights clinically important predicates. The present implementation approximates this with phrase embeddings and abnormality support scores. Predicate-level modeling is a natural extension because some explanations contain both grounded and ungrounded parts. For example, a sentence may correctly mention a visible opacity but incorrectly assign its location. A sentence-level score can blur this distinction.

The relation between answer correctness and grounding can be represented as a joint distribution. Let A be correctness and H groundedness. Four output types exist: correct-grounded, correct-ungrounded, incorrect-grounded, and incorrect-ungrounded. The ordinary accuracy metric estimates $\Pr(A = 1)$. A grounding-aware metric estimates $\Pr(A = 1, H = 1)$. The gap is

$$\Pr(A = 1) - \Pr(A = 1, H = 1) = \Pr(A = 1, H = 0). \tag{17}$$

In the test set, this term was large. It represents correct answers whose explanations failed grounding criteria. These are not necessarily clinically wrong answers, but they are weak explanations. The distinction is important when models are used for teaching, justification, or user trust.

Table 8: Negative-Control and Mask Sensitivity Analysis

Analysis	Observation	Rate/Effect	Interpretation
Negative-control instability	False positives after edits	6.8%	Artifact sensitivity present
Background-edit instability	Incorrect positive conversion	5.9%	General edit response
Dilated-mask erasure	Collapse increase	+7.4%	More evidence removed
Contracted-mask erasure	Collapse decrease	-11.6%	Incomplete evidence removal
Grounded Λ_i median	Asymmetry ratio	2.41	Lesion-specific change
Ungrounded Λ_i median	Asymmetry ratio	1.08	Non-specific edit response

Table 9: Major Failure Modes Identified

Failure Mode	Description	Typical Behavior	Most Affected Area
Language inertia	Repeated explanations	Same wording after edits	Pleural findings
Edit overreactivity	Changes after any edit	Unstable explanations	Dermatology images
Generic reasoning	Non-specific evidence	"Image supports it"	Compact models
Relocation inconsistency	Wrong spatial interpretation	Old location reused	Border irregularity
Weak answer-explanation coupling	Mismatched updates	Answer fixed, reason altered	Lower-ranked models

The relocation condition can be represented as a test of equivariance. If a visual feature moves from location ℓ_1 to location ℓ_2, a grounded explanation should update location-dependent language. Let $T_{\ell_1 \to \ell_2}$ be the relocation transformation. Let G be the explanation generator. A location-equivariant explanation would satisfy

$$G\left(T_{\ell_1 \to \ell_2} x, q\right) \approx T^{text}_{\ell_1 \to \ell_2} G(x, q), \tag{18}$$

where T^{text} changes the explanation's location terms appropriately. Relocation contradiction occurs when the left side remains close to $G(x, q)$ instead of the transformed explanation. This is why similarity to the original output becomes a negative signal in R_i when anatomical consistency is low.

The low-rank residual analysis provides another view of model behavior. Let each output have a vector of counterfactual responses $y_i = (C_i, B_i, R_i)$. The matrix decomposition

$$Y = XW + PQ^\top + E \tag{19}$$

separates finding-type expectations from residual behavior. The first residual component in P was high for outputs that repeated explanations across all edits. The second was high for outputs that changed explanations after both lesion and background erasure. These two components correspond to two distinct failure modes: language inertia and edit overreactivity. A model can have one without the other. For example, Model Five had high language inertia, while Model Four had high edit overreactivity.

The grounding coefficient can be interpreted as a monotone function of three desirable inequalities. A well-grounded explanation should satisfy

$$\begin{aligned} s(e_i^0) &> s(e_i^-), \\ \mathrm{sim}(e_i^0, e_i^b) &> \mathrm{sim}(e_i^0, e_i^-), \\ \mathrm{cons}(e_i^r, x_i^r, q_i) &> \mathrm{sim}(e_i^0, e_i^r) - \delta. \end{aligned} \tag{20}$$

The first inequality says lesion erasure should reduce abnormality support. The second says background erasure should disturb the explanation less than lesion erasure. The third says relocation should preserve consistency rather than merely preserve the original wording. The learned coefficient softens these inequalities and weights them according to empirical expert labels.

One might ask why explanation collapse should be required if the model's answer remains correct after erasure. In some cases, erasure may fail to remove all evidence, or the abnormality may have distributed signs outside the mask. The framework addresses this by using expert masks, dilation sensitivity, and background controls, but it cannot eliminate all ambiguity. Therefore, collapse is not treated as an absolute proof. It is a probabilistic signal. The latent coefficient combines multiple signals and allows partial evidence. This is more appropriate than declaring every non-collapsing explanation ungrounded.

Another mathematical issue is measurement error from edited images. Let Q_i be an edit-quality variable indicating whether inpainting successfully removes evidence without artifacts. If Q_i is low, collapse metrics become unreliable. The ideal model would condition on Q_i:

$$\Pr(h_i = 1) = \sigma\left(\beta^\top z_i + \eta Q_i + \beta_Q^\top (z_i Q_i)\right). \tag{21}$$

The current study approximates this by quality-control filtering and by including background stability. A future version could explicitly estimate edit quality for every case and downweight uncertain interventions.

7 Discussion

The study shows that answer correctness and evidence-grounded explanation are sharply different measurements. Correct answers were common, but counterfactually grounded correct explanations were much less common. This does not mean that most correct answers were useless. It means that many explanations did not behave as though they depended on the image region they described. In clinical education, audit, or user-facing explanation, that distinction matters. A correct answer with an ungrounded explanation may create misplaced confidence in the model's visual reasoning.

Counterfactual lesion erasure offers a practical way to test explanation dependence without needing access to model internals. It asks the model the same question after controlled changes to the image. If the explanation remains unchanged after removal of the evidence, the explanation is suspect. If it changes after irrelevant background removal, the model may be sensitive to editing artifacts. If it repeats the original explanation after evidence relocation, it may not understand the spatial context of its own claim. These behavioral tests are more directly tied to output faithfulness than raw attention maps.

The results also caution against relying on attention overlap as a grounding proxy. Mask-language alignment improved over correctness alone, but it did not match counterfactual scoring. A model may align a phrase with the right region because the region is visually salient, yet its generated explanation may not depend on that region. Conversely, some explanations may be grounded in distributed evidence not captured by a tight mask. Attention and alignment remain useful diagnostic tools, but they should be paired with interventions that alter the evidence.

Different findings produced different grounding behavior [13]. Device and hardware explanations were most grounded, likely because their visual features are high contrast and spatially definite. Pleural abnormalities were least grounded, likely because their signs are subtle and language templates are common. Dermatologic lesion explanations showed strong erasure response but also more background instability. This suggests that grounding evaluation should be finding-specific. A single aggregate score may hide important weaknesses in subtle or texture-based findings.

Relocation was one of the more revealing interventions. Some models responded appropriately to erasure but failed after relocation. This means they could detect absence but not reinterpret spatial meaning. Medical explanation often depends on spatial relations. A feature is not only present; it is present in a particular anatomical location. A fracture-like line outside the bone should not support a fracture explanation. A border irregularity moved away from the lesion should not support the same lesion-boundary explanation. Relocation tests this spatial relation more directly than erasure [14].

Generic explanations were common. They were not always wrong, but they were not informative. In medical settings, an explanation should provide more than a restatement of the answer. If a model says the abnormality is present because it is visible, it has not shown what visual evidence supports the claim. Generic explanations also tend to evade counterfactual collapse because they contain little evidence-specific language [15]. A model can appear stable simply because it says almost nothing. Evaluation should therefore penalize generic reasons when explanation is requested.

The method has implications for dataset design [16]. To evaluate explanation grounding, reference labels alone are insufficient. Expert masks or region-level evidence annotations are needed. These annotations need not be pixel-perfect, but they must identify the region whose removal should affect the explanation. Creating such masks is labor-intensive, yet the value is substantial because it enables causal-style tests. Future medical VLM benchmarks should consider adding evidence masks for a subset of cases rather than only image-level labels and report text.

The method also has implications for model development. Training models to produce explanations from reports may encourage plausible language without visual dependence. A report phrase can be learned as a common continuation for a finding. Counterfactual training could penalize explanations that persist after evidence removal. For example, a model could be trained with original and lesion-erased pairs so that abnormality-specific explanation appears only when evidence is present [17]. This would require careful editing quality control, but the present results suggest it could improve faithfulness.

One should not interpret failure to collapse as definitive proof that the model ignored the image. There are several alternative explanations. The erasure might not remove all relevant evidence. The abnormality might have secondary signs outside the mask. The model might rely on global context that remains valid. The explanation parser might miss subtle changes. This is why the paper uses multiple metrics, human labels, and mask sensitivity

analysis. Counterfactual evaluation is strongest when erasure, background control, relocation, and expert review agree.

The benchmark's negative controls are important because edited images can create artifacts [18]. If a model changes after any inpainting, collapse after lesion erasure is not meaningful. Background-erasure stability helps identify this problem. Dermatology images were more sensitive to background edits, reminding us that image editing is not neutral. Counterfactual tests must verify that the intervention changes the intended evidence rather than introducing an artifact-driven shortcut. This is particularly important in modalities with texture, color, or lighting variation.

The difference between language inertia and edit overreactivity is useful for diagnosing model behavior [19]. Language inertia means the model repeats explanation patterns despite image changes. Edit overreactivity means the model changes output whenever the image is altered, even outside the evidence region. Both are failures, but they require different remedies. Language inertia may require stronger visual grounding objectives. Edit overreactivity may require robustness to benign image variation and better artifact handling. The low-rank analysis separated these behaviors across models.

A practical deployment question is whether every model output should undergo counterfactual testing. The answer is probably no. Generating three edited variants for every clinical image would be expensive and may not be necessary. Counterfactual testing is better suited for model evaluation, development, periodic audit, and high-value explanation validation [20]. A smaller audit set with expert masks can reveal whether a model's explanations are generally trustworthy for certain findings. For real-time use, a lighter predictor trained from counterfactual audits may flag outputs likely to be ungrounded.

The results also raise a communication issue. If a model's explanation is not grounded, it may be better to produce no explanation than a misleading one. Users often interpret explanations as evidence of reasoning. A fluent but ungrounded reason can increase trust without justification. Systems should therefore distinguish between answer mode and explanation mode. A model may be allowed to answer a simple question but not allowed to provide a visual explanation unless it can meet grounding criteria. This separation may be important in patient-facing or educational tools.

The study supports a shift from explanation appearance to explanation behavior. A sentence that resembles a radiology rationale is not necessarily a faithful rationale. Behavior under counterfactual evidence changes provides a stronger test. This aligns with a broader principle in medical artificial intelligence: outputs should be evaluated not only by what they say on unmodified examples, but by how they respond when clinically relevant evidence is changed in controlled ways.

8 Limitations

The first limitation is the use of edited images. Inpainting approximates evidence removal but cannot perfectly create the true counterfactual image that would have existed without the lesion. Some edits may leave residual evidence, and others may introduce artifacts. The study used quality control, background controls, and mask sensitivity analysis to reduce this problem, but no editing method is perfect. The counterfactual should be interpreted as a probe, not as a literal alternate clinical reality.

The second limitation is mask dependence. Expert masks vary across annotators, and some findings have diffuse evidence. Pleural effusion, edema, and inflammatory skin change can extend beyond a compact region. If the mask is too narrow, erasure may not remove enough evidence. If it is too broad, it may remove contextual anatomy. The dilation and contraction analysis showed that mask size affects collapse. Future work should use probabilistic evidence maps rather than binary masks for diffuse findings.

The third limitation concerns explanation parsing. The output parser separated answers from explanations and extracted visual phrases, but parsing errors remain. Some models write explanations in compressed or unusual forms. Others combine answer and evidence in one phrase. Parser mistakes can distort collapse and alignment metrics. Human audit reduced this problem in the labeled subset, but full automation still depends on reliable medical language parsing. Predicate-level parsing would likely improve the analysis.

The fourth limitation is modality coverage. The benchmark used radiography and dermatology. It did not include CT, MRI, ultrasound video, pathology whole-slide images, endoscopy, ophthalmology, or nuclear medicine. These modalities have different evidence structures. For example, CT findings can span slices, pathology evidence may require magnification changes, and ultrasound depends on dynamic scanning. Counterfactual erasure would need adaptation for each modality. The current findings should not be treated as universal.

The fifth limitation is model anonymity. The systems are named generically to keep the analysis focused on behavior rather than ranking. This limits direct reproducibility for readers who want to compare named systems. The tradeoff was chosen because model interfaces and versions change, while the evaluation principle is more stable. Future open benchmarking could report named open models under fixed version control.

The sixth limitation is that the prompt asked for one short reason. Longer explanations might show different grounding behavior. A model may provide a generic reason when constrained to be brief but a more specific reason when allowed longer output. Conversely, longer output may introduce more ungrounded details. The present paper evaluates compact explanation because it is common in question-answering settings [21]. Full report rationales require separate study.

The seventh limitation is that expert groundedness labels are themselves judgment-based [22]. Reviewers can disagree about whether a phrase such as increased opacity at the base is sufficiently tied to a mask. Some explanations refer to distributed patterns rather than a single region [23]. The study used consensus labels and measured agreement, but the target remains partly subjective. Better annotation interfaces that show phrase-region links may improve reliability.

The eighth limitation is computational cost. Each case required multiple edited images and repeated model queries. This is feasible for research evaluation but heavy for routine use. Future work should identify smaller counterfactual audit sets that predict broader explanation behavior. Active sampling may reduce cost by selecting cases where grounding is uncertain or clinically important.

The ninth limitation is that the method tests local evidence dependence, not complete reasoning. A model may pass lesion-erasure tests while still making poor clinical judgments in complex cases. Conversely, a model may fail a grounding test for a subtle explanation but still provide a useful answer. Counterfactual grounding is one component of evaluation. It should be combined with diagnostic accuracy, uncertainty assessment, safety review, and clinical workflow testing.

The tenth limitation is the reliance on deterministic querying. Some models may vary under different decoding settings. The study used deterministic settings to reduce noise and isolate image-condition effects. In real applications, stochastic decoding or model updates may change explanations. Future work should measure grounding stability across decoding seeds and model versions.

9 Future Work

Future research should move toward predicate-level counterfactual grounding. Instead of scoring an entire explanation sentence, the system should decompose it into claims such as location, sign, severity, and visual descriptor. Each predicate should be tested against evidence. This would handle mixed explanations where one phrase is grounded and another is not. It would also allow more precise feedback to model developers.

Another direction is counterfactual training. Models could be trained with paired original and lesion-erased images so that explanations align with evidence presence. Such training must be careful because inpainting artifacts can become shortcuts. A good training procedure would include background erasure, relocation, mask uncertainty, and multiple editing methods [24]. The objective should reward lesion-specific response rather than edit detection.

A third direction is extension to volumetric imaging. For CT and MRI, evidence often spans slices and planes. Counterfactual erasure would need three-dimensional masks and slice-consistent editing. Explanation collapse would need to consider whether the model refers to the correct slice, organ, and spatial relation. This could be especially useful for findings such as pulmonary embolism, liver lesions, hemorrhage, or spinal stenosis.

A fourth direction is human evaluation of explanation usefulness. The present paper evaluates whether explanations are grounded, not whether they help clinicians or patients. A grounded explanation may still be too technical, too vague, or unnecessary. Human studies could test whether counterfactually grounded explanations improve error detection, learning, or appropriate trust. Such studies should compare grounded explanations, generic explanations, and no explanations.

A fifth direction is real-time grounding prediction. Since full counterfactual testing is expensive, a lightweight predictor could estimate grounding risk from the original image, question, and response. The present study showed that original-output-only prediction is weaker than counterfactual evaluation, but it may still be useful when trained on counterfactual audit data. Such a predictor could flag explanations that should be withheld or reviewed [25].

A sixth direction is multi-mask evaluation. Some findings have several evidence regions. A fracture may have cortical break, soft tissue swelling, and alignment change. A melanoma-like lesion may have border irregularity, color variation, and asymmetry. Erasing one region may not collapse the explanation if another remains. Multi-mask testing could estimate which evidence components support which explanation predicates.

A seventh direction is comparison of editing methods. The present study used modality-specific inpainting and patch synthesis. Other approaches include generative counterfactuals, diffusion-based editing, feature-space erasure, and acquisition-aware simulation. Each method has different artifacts and assumptions [26]. A robust evaluation should check whether grounding conclusions remain stable across editing methods.

An eighth direction is integration with reporting interfaces. If a model provides an explanation, the interface could highlight the region that supports each phrase. If no reliable region can be identified, the explanation could be suppressed or marked as uncertain. Counterfactual evaluation could be used offline to validate such highlighting systems. This would connect evidence grounding to user experience.

A ninth direction is evaluation under distribution shift. Grounding may change when images come from different scanners, institutions, patient populations, or acquisition protocols. A model may rely more on language priors when image quality is unfamiliar. Counterfactual tests across institutions could reveal whether grounding is stable or dataset-specific. This is particularly important for rare findings.

A tenth direction is combining counterfactual grounding with uncertainty. If lesion erasure removes evidence, a model might not need to reverse its answer completely; it might appropriately become uncertain. Future metrics should distinguish uncertainty increase from wrong persistence. A model that says the finding is no longer clearly visible after erasure is behaving more appropriately than one that confidently repeats the original explanation. Groundedness and uncertainty should therefore be modeled together.

10 Conclusion

This paper introduced a counterfactual lesion-erasure framework for measuring visual grounding in medical vision-language explanations. The method uses expert-marked evidence regions, lesion erasure, background erasure, and evidence relocation to test whether generated explanatory language changes when the visual basis for that language is altered. The evaluation focuses on the relation between a stated reason and the image region that should support it.

The empirical results showed a substantial gap between answer correctness and grounded explanation. Across radiographic and dermatologic cases, models often answered correctly while their explanations persisted after the relevant evidence was removed. Some models reacted to irrelevant image edits, while others repeated language patterns across all conditions. Relocation revealed additional failures in spatial consistency that erasure alone did not detect. These findings indicate that fluent medical explanations should not be assumed faithful merely because the answer is correct.

The latent grounding coefficient improved prediction of expert-labeled groundedness by combining collapse, background stability, relocation contradiction, mask-language alignment, correctness, and model-domain effects. Counterfactual metrics outperformed answer correctness and static alignment alone. The results support the use of controlled image interventions as a stronger test of explanation behavior than ordinary output inspection.

The framework has limitations. Edited images are imperfect counterfactuals, masks can be incomplete, and explanation parsing remains difficult. The benchmark covers radiography and dermatology rather than all medical imaging. The method is also too expensive for routine real-time use in its full form. Even so, it provides a concrete way to audit whether generated explanations depend on the evidence they claim to describe.

Medical vision-language systems should be evaluated not only on what they answer, but also on whether their reasons follow the image. A useful explanation should weaken when its evidence disappears, remain stable when irrelevant background changes, and update when evidence moves. Counterfactual lesion erasure makes these expectations measurable. It offers a path toward medical explanations that are not merely plausible sentences but evidence-responsive statements.

References

[1] J. L. Moe, G. Pappas, and A. Murray, "Transformational leadership, transnational culture and political competence in globalizing health care services: A case study of jordan's king hussein cancer center," *Globalization and health*, vol. 3, no. 1, pp. 11–11, Nov. 16, 2007. DOI: 10.1186/1744-8603-3-11

[2] J. C. McBee et al., "Assessing chatgpt's competency in addressing interdisciplinary inquiries on chatbot uses in sports rehabilitation: Simulation study.," *JMIR medical education*, vol. 10, e51157–e51157, Aug. 7, 2024. DOI: 10.2196/51157

[3] B. M. Scott et al., "Post-exercise pulse pressure is a better predictor of executive function than pre-exercise pulse pressure in cognitively normal older adults.," *Neuropsychology, development, and cognition. Section B, Aging, neuropsychology and cognition*, vol. 23, no. 4, pp. 464–476, Dec. 2, 2015. DOI: 10.1080/13825585.2015.1118007

[4] K. K. Sudnawa et al., "Heterogeneity of comprehensive clinical phenotype and longitudinal adaptive function and correlation with computational predictions of severity of missense genotypes in kif1a-associated neurological disorder.," *Genetics in medicine : official journal of the American College of Medical Genetics*, vol. 26, no. 8, pp. 101 169–101 169, May 21, 2024. DOI: 10.1016/j.gim.2024.101169

[5] B. Sadanandan and V. Behzadan, "Vsf-med: A vulnerability scoring framework for medical vision-language models," *arXiv preprint arXiv:2507.00052*, 2025.

[6] J. A. Deal et al., "Hearing treatment for reducing cognitive decline: Design and methods of the aging and cognitive health evaluation in elders randomized controlled trial.," *Alzheimer's & dementia (New York, N. Y.)*, vol. 4, no. 1, pp. 499–507, Oct. 5, 2018. DOI: 10.1016/j.trci.2018.08.007

[7] D. G. Morrow et al., "An emr-based tool to support collaborative planning for medication use among adults with diabetes: Design of a multi-site randomized control trial," *Contemporary clinical trials*, vol. 33, no. 5, pp. 1023–1032, Jun. 1, 2012. DOI: 10.1016/j.cct.2012.05.010

[8] L. Fleming, C. Sheridan, D. Waite, M. G. Klug, and L. Burd, "Screening for fetal alcohol spectrum disorder in infants and young children.," *Advances in drug and alcohol research*, vol. 3, pp. 11 125–11 125, Jul. 14, 2023. DOI: 10.3389/adar.2023.11125

[9] D. R. Perszyk, B. Ferguson, and S. R. Waxman, "Maturation constrains the effect of exposure in linking language and thought: Evidence from healthy preterm infants," *Developmental science*, vol. 21, no. 2, Dec. 29, 2016. DOI: 10.1111/desc.12522

[10] J. Applequist, M. Miller-Day, P. F. Cronholm, R. A. Gabbay, and D. S. Bowen, ""in principle we have agreement, but in practice it is a bit more difficult": Obtaining organizational buy-in to patient-centered medical home transformation.," *Qualitative health research*, vol. 27, no. 6, pp. 909–922, Nov. 30, 2016. DOI: 10.1177/1049732316680601

[11] C. J. Peek, M. L. Allen, J. T. Pacala, W. Nickerson, and A. Westby, "Coming together in action for equity, diversity, and inclusion.," *Family medicine*, vol. 53, no. 9, pp. 786–795, Jul. 21, 2021. DOI: 10.22454/fammed.2021.569762

[12] V. T. Lai and L. Boroditsky, "The immediate and chronic influence of spatio-temporal metaphors on the mental representations of time in english, mandarin, and mandarin-english speakers," *Frontiers in psychology*, vol. 4, pp. 142–142, Jul. 10, 2013. DOI: 10.3389/fpsyg.2013.00412;10.3389/fpsyg.2013.00142

[13] V. Manivannan, "What we see when we digitize pain: The risk of valorizing image-based representations of fibromyalgia over body and bodily experience.," *Digital health*, vol. 3, pp. 2 055 207 617 708 860–, May 22, 2017. DOI: 10.1177/2055207617708860

[14] J. A. Rich, "Viewing the future through the lens of the past: A personal reflection on disparities education in medicine and public health," *Journal of general internal medicine*, vol. 25, no. 2, pp. 202–203, Mar. 30, 2010. DOI: 10.1007/s11606-010-1302-4

[15] P. G. Williams et al., "The pediatrician's role in optimizing school readiness," *Pediatrics*, vol. 138, no. 3, Sep. 1, 2016. DOI: 10.1542/peds.2016-2293

[16] B. D. Kelly et al., "Radiology artificial intelligence, a systematic evaluation of methods (raise): A systematic review protocol.," *Insights into imaging*, vol. 11, no. 1, pp. 1–6, Dec. 9, 2020. DOI: 10.1186/s13244-020-00929-9

[17] J. R. Gimpel et al., "Ugrc 2021 recommendations on gme transition: Pros and cons, opportunities and limitations.," *Journal of osteopathic medicine*, vol. 122, no. 9, pp. 461–464, May 12, 2022. DOI: 10.1515/jom-2021-0285

[18] T. A. Koleck, C. Dreisbach, P. E. Bourne, and S. Bakken, "Natural language processing of symptoms documented in free-text narratives of electronic health records: A systematic review," *Journal of the American Medical Informatics Association : JAMIA*, vol. 26, no. 4, pp. 364–379, Feb. 6, 2019. DOI: 10.1093/jamia/ocy173

[19] C. M. Startin et al., "Health comorbidities and cognitive abilities across the lifespan in down syndrome.," *Journal of neurodevelopmental disorders*, vol. 12, no. 1, pp. 4–4, Jan. 23, 2020. DOI: 10.1186/s11689-019-9306-9

[20] G. Wu, D. A. Lee, W. Zhao, A. Wong, and S. Sidhu, "Chatgpt: Is it good for our glaucoma patients?" *Frontiers in ophthalmology*, vol. 3, pp. 1 260 415–1 260 415, Nov. 16, 2023. DOI: 10.3389/fopht.2023.1260415

[21] R. D. Boyce et al., "Bridging islands of information to establish an integrated knowledge base of drugs and health outcomes of interest," *Drug safety*, vol. 37, no. 8, pp. 557–567, Jul. 2, 2014. DOI: 10.1007/s40264-014-0189-0

[22] J. Cheng, "Applications of large language models in pathology.," *Bioengineering (Basel, Switzerland)*, vol. 11, no. 4, pp. 342–342, Mar. 31, 2024. DOI: 10.3390/bioengineering11040342

[23] A. Schmaltz and A. L. Beam, "Sharpening the resolution on data matters: A brief roadmap for understanding deep learning for medical data," *The spine journal : official journal of the North American Spine Society*, vol. 21, no. 10, pp. 1606–1609, Aug. 25, 2020. DOI: 10.1016/j.spinee.2020.08.012

[24] G. Hu and B. Yu, "Artificial intelligence and applications," *Journal of Artificial Intelligence and Technology*, vol. 2, no. 2, pp. 39–41, Apr. 5, 2022. DOI: 10.37965/jait.2022.0102

[25] J. Wolfer, "Embedding topical elements of parallel programming, computer graphics, and artificial intelligence across the undergraduate cs required courses," *International Journal of Engineering Pedagogy (iJEP)*, vol. 5, no. 1, pp. 27–32, Feb. 11, 2015. DOI: 10.3991/ijep.v5i1.4090

[26] K. Gottlieb, "The nuka system of care: Improving health through ownership and relationships," *International journal of circumpolar health*, vol. 72, no. 1, pp. 21 118–, Jan. 31, 2013. DOI: 10.3402/ijch.v72i0.21118