CORRUPTION-RISK REVERSAL IN ADAPTIVE VISUAL EVIDENCE ROUTING FOR MEDICAL VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adaptive visual computation is usually motivated as an efficiency problem: additional visual tokens are spent only when a global view is insufficient. For an autoregressive vision–language model (VLM), however, an acquired crop is not simply more information. It changes both the visible evidence and the generation context, and can therefore overturn a correct global answer. This reliability problem is formulated as cost-sensitive evidence routing among answering, acquiring a view, and abstaining. The formulation separates the availability of a better action from the router's ability to select it, and decomposes acquisition value into helpful corrections and harmful corruptions. Empirically, every policy is reconstructed from immutable paired counterfactual outputs with a shared global response, exact-image grouping, revision-pinned models, and policies frozen before transfer. On historically exposed M3CoTBench development, signed-utility routing raises MedGemma-4B accuracy from 24.55% to 30.91% ($+6.36$ pp; 95% image-group bootstrap CI $[+2.59, +11.21]$). Under locked PathVQA shift, the same rule lowers accuracy from 70.31% to 63.67% (-6.64 pp; CI $[-10.55, -2.73]$). A separately frozen confidence-ranked plurality policy likewise lowers Qwen accuracy on 2,996 prospective 3MDBench cases from 40.92% to 39.05%, correcting 16 answers and corrupting 72. On DDI, a conservative unanimity gate avoids added errors only by making no replacements, while global accuracy conceals 0.58% malignant sensitivity. Across these evaluations, the observed pattern is termed *corruption-risk reversal*: uncertainty continues to identify fragile cases, but ceases to distinguish acquisitions that repair an answer from those that damage it. These results establish neither a universal failure of cropping nor clinical safety. They instead show that efficiency claims alone are insufficient without action headroom, transferable harm estimation, safe aggregation, and nontrivial coverage.

1 INTRODUCTION

A medical image rarely has a single useful scale. A global view preserves anatomy and spatial context, whereas a focused view can expose a small lesion, cellular texture, or device boundary. This motivates systems that allocate visual computation adaptively: answer easy questions from the full image and spend extra tokens, crops, or tool calls on difficult ones (Ryoo et al., 2021; Lin et al., 2025; Carvalho et al., 2025). The usual justification is computational. If targeted evidence is more informative than the global representation, selective acquisition should approach the accuracy of expensive multi-view inference at lower average cost.

That argument moves too quickly from information to decision quality. For a fixed statistical observer, revealing an additional measurement enlarges the available information set. For an autoregressive VLM, a crop also changes the input distribution, suppresses global context, alters attention and token competition, and may activate a different language prior. The acquired response is therefore a new model action, not a guaranteed refinement of the old one. A crop can correct a wrong answer, leave it unchanged, or corrupt an answer that was already correct. The last event is obscured when studies report only average accuracy, acquisition rate, or oracle localization quality. It is also poorly captured by self-reported confidence: calibration of current correctness does not imply calibration of the value of changing the evidence (Guo et al., 2017).

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

Figure 1 operationalizes this distinction as three linked audits. During paired counterfactual development, every candidate action is compared with the same global decode, so its signed outcome records whether acquisition corrects an error, leaves the decision unchanged, or corrupts a previously correct answer. At test time, the router receives only pre-acquisition variables and must choose among retaining the global response, acquiring one view, and abstaining before any crop response exists. Selection and aggregation are then evaluated as a separate stage rather than being credited with oracle access to the acquired outputs. Finally, the complete policy is frozen before target-domain evaluation. This organization separates three questions that aggregate accuracy conflates: whether the action set contains a better answer, whether the router can identify it without post-acquisition leakage, and whether the answer-selection rule preserves the available gain after transfer.

The empirical sequence is deliberately asymmetric. M3CoTBench development is historically exposed and is used only to construct and understand the policy. The resulting rule is then held fixed under model and PathVQA transfer. Two subsequent tests are frozen prospectively: a confidence-ranked multi-view policy on 3MDBench and a conservative unanimity gate on DDI. The central observation is not that crops are intrinsically harmful. Rather, the sign and ranking of their conditional utility can change. On MedGemma, help remains partly rankable across M3CoTBench and PathVQA while harm ranking reverses; on 3MDBench, consensus among additional views amplifies correlated errors; on DDI, preventing corruption through a strict gate produces zero useful replacements and leaves near-zero malignant sensitivity untouched.

This empirical pattern is termed *corruption-risk reversal*. The phrase names a measured failure mode, not a distribution-free theorem: a policy may continue to find cases likely to change while losing the ability to tell beneficial changes from destructive ones. The paper contributes a finite-horizon decision formulation with answer and abstention options, an exact help–harm identity and action-headroom decomposition, an auditable paired protocol with locked transfer, and a set of negative results that localize failure to action availability, selection, or aggregation. The practical implication is correspondingly limited but important: an adaptive visual system should report not only accuracy and compute, but also recoverable headroom, help and harm separately, aggregation regret, and coverage.

2 Related Work

Adaptive computation in vision reduces redundant processing by selecting tokens, regions, or resolutions as a function of the input. TokenLearner compresses spatial structure into a small set of learned adaptive tokens (Ryoo et al., 2021); more recent VLM systems expose cropping as an explicit tool. AdaptVision learns when to invoke a bounding-box acquisition, whereas CropVLM trains a zoom module that can be attached to a frozen VLM (Lin et al., 2025; Carvalho et al., 2025). These methods establish that conditional visual processing can improve an accuracy–cost trade-off. The unresolved question is whether replacing or aggregating the original answer remains reliable after a new view is acquired. The distinction matters because token selection inside a discriminative network and a new autoregressive decode after cropping need not preserve the same information or error structure.

The decision problem is related to active feature acquisition, Bayesian experimental design, and optimal stopping. EDDI selects costly measurements by expected information gain under a partial variational model (Ma et al., 2019); acquisition-conditioned oracles and more recent ranking approaches seek useful features without enumerating a full generative model (Valancius et al., 2024; Guney et al., 2025). These frameworks motivate trading predictive value against cost, while classical stopping theory supplies the recursion for deciding whether to stop or continue (Chow, 1970; Chow et al., 1971). Selective prediction adds a reject action and evaluates the resulting risk–coverage trade-off (Geifman & El-Yaniv, 2019); conformal risk control can provide finite-sample guarantees for suitable monotone losses under exchangeability (Angelopoulos et al., 2024). The present setting combines these elements but introduces a particular obstacle: the acquisition acts through a generative VLM, and its utility is signed. Estimating that the answer is uncertain, or that the crop is informative, does not determine whether a new decode will correct or corrupt the current response.

This obstacle is consistent with broader evidence that VLM outputs can follow linguistic priors rather than image content. POPE documents object hallucination in large VLMs, and HallusionBench separates language hallucination from visual illusion through controlled question pairs (Li et al., 2023; Guan et al., 2024). Medical evaluation adds domain shift, class imbalance, and clinically asym-

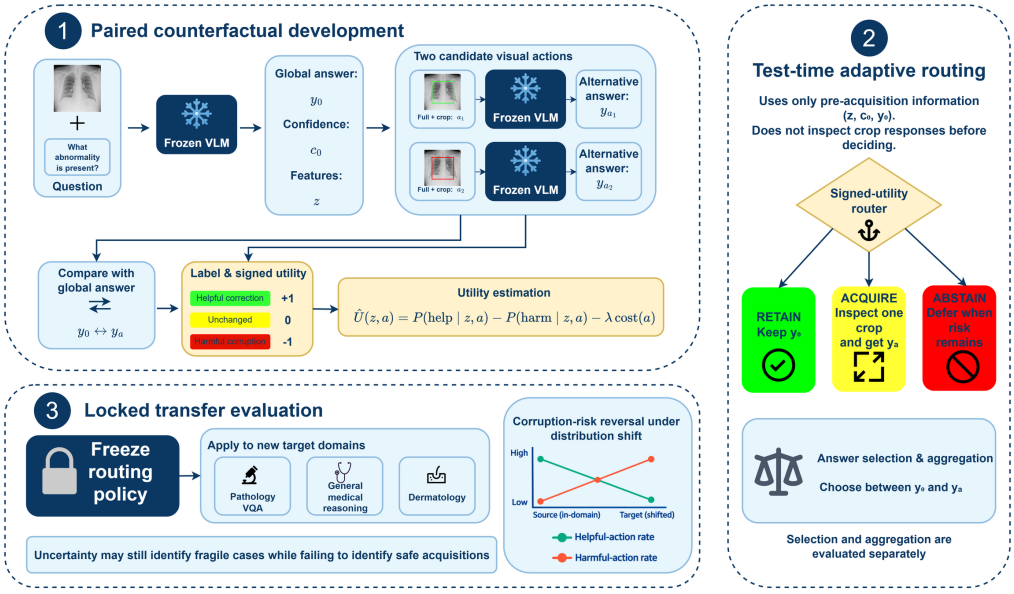


Figure 1: Study design and inferential boundary. Paired development uses a frozen VLM and a shared global answer y_0 to label each visual action as a helpful correction, an unchanged decision, or a harmful corruption, yielding a signed utility target. The deployed router may use only (z, c_0, y_0) when choosing to retain, acquire, or abstain; crop responses remain unavailable until after acquisition, and subsequent answer selection is audited separately. The policy is then locked for target-domain evaluation, where opposing shifts in helpful and harmful action rates constitute the measured corruption-risk reversal.

metric errors. PathVQA and SLAKE provide pathology and semantically annotated medical VQA settings (He et al., 2021; Liu et al., 2021); M3CoTBench and 3MDBench broaden multimodal reasoning and diagnosis protocols (Jiang et al., 2026; Sviridov et al., 2025). Generalist biomedical systems demonstrate the breadth now possible (Zhang et al., 2024), but strong benchmark performance does not by itself establish reliability in clinical decision workflows (Hager et al., 2024). The evaluation includes Qwen2.5-VL and MedGemma as open general and medically tuned models (Bai et al., 2025; Sellergren et al., 2025), and uses DDI's pathology-confirmed dermatology images for an external class-conditional stress test (Daneshjou et al., 2022). This study does not rank these models against the broader medical-VLM literature; it asks how the same model's answer changes under controlled evidence interventions and whether those changes remain predictable after transfer.

3 Decision Formulation

Let h_t contain all information observed before decision t: the global image x, question q, previous evidence and model outputs. Let $\mathcal{A}(h_t)$ contain unobserved acquisition actions, $L_{\text{ans}}(h_t)$ the loss of answering now, α the abstention loss, and $c(h_t, a)$ the acquisition cost. The valid Bellman recurrence is

$$V_T(h_T) = \min\{L_{\text{ans}}(h_T), \alpha\}, \tag{1}$$

$$V_t(h_t) = \min\left\{L_{\text{ans}}(h_t), \alpha, \min_{a \in \mathcal{A}(h_t)} [c(h_t, a) + \mathbb{E}\{V_{t+1}(h_{t+1}) \mid h_t, a\}]\right\}. \tag{2}$$

The acquisition decision must be measurable with respect to h_t; it cannot inspect a crop response before deciding to buy it. Equation 2 also shows why a raw confidence-gain rule is not generally optimal when abstention is available: the correct comparison is against the smaller of answer and abstention losses.

The recurrence clarifies the decision boundary, but the logged experiments expose one-step counterfactual actions rather than the full continuation tree. The analysis therefore turns to quantities

that are exactly identifiable from the paired cache. For a one-step candidate a, let $C_0, C_a \in \{0, 1\}$ denote correctness of the shared global response and the acquired response. Define

$$G_a = C_a - C_0 \in \{-1, 0, 1\}, \tag{3}$$

$$U_a(z) = P(G_a = 1 \mid z, a) - P(G_a = -1 \mid z, a) - \lambda_a, \tag{4}$$

where z contains only pre-acquisition features and λ_a is normalized cost. The router estimates separate help and harm probabilities and applies

$$a^* = \arg\max_a \widehat{U}_a(z), \qquad \text{acquire iff } \widehat{U}_{a^*}(z) > 0. \tag{5}$$

Proposition 1 (Help–harm identity)**.** *For any fixed action a and data distribution D,*

$$\mathbb{E}_D[C_a - C_0] = P_D(G_a = 1) - P_D(G_a = -1). \tag{6}$$

Thus an acquisition improves accuracy only when its correction probability exceeds its corruption probability; unchanged cases cancel.

Proof. Expand $\mathbb{E}[G_a]$ over its three support values $1, 0, -1$. □

This elementary identity motivates the estimator but is not itself a novel result. The empirical question is whether both terms transfer. For a policy π that returns a chosen or aggregated prediction C_π, define the restricted action oracle $C^* = \max(C_0, \max_a C_a)$. Then

$$\underbrace{\mathbb{E}[C_\pi - C_0]}_{\text{realized gain}} = \underbrace{\mathbb{E}[C^* - C_0]}_{H_D \text{ (action headroom)}} - \underbrace{\mathbb{E}[C^* - C_\pi]}_{R_D(\pi) \text{ (selection/aggregation regret)}}. \tag{7}$$

Equation 7 separates absence of useful actions from failure to select or combine them.

Proposition 2 (Conditional action-value bound)**.** *If every available action value $Q_t(h, a)$, including answer and abstain, has an estimate satisfying $|\widehat{Q}_t - Q_t| \le \epsilon_t$, then the action minimizing $\widehat{Q}_t$ has one-step regret at most $2\epsilon_t$. If the estimates are optimal continuation values at every reachable history, total finite-horizon excess loss is at most $2\sum_t \epsilon_t$.*

Proof. Let a^* minimize Q_t. Add and subtract $\widehat{Q}_t$ for the chosen action and a^*. Estimated optimality removes the middle term and the two remaining errors are bounded by ϵ_t. Backward induction gives the horizon bound. □

No such uniform bound is established for the fitted router; the proposition clarifies the quantity that would need to transfer.

The same accounting extends from selection to aggregation. Let f_k aggregate the global response with k acquired views and define realized marginal value

$$\delta_k = \mathbb{E}\left[C(f_k) - C(f_{k-1})\right]. \tag{8}$$

A monotone-information intuition expects nonnegative, often diminishing δ_k. Autoregressive answers need not satisfy it: new views can change the label, and correlated errors can dominate a vote. Equation 8 is tested over all 24 orders of four quadrant views.

4 Experimental Design and Audit Protocol

The evaluation is organized by inferential role rather than by dataset alone. This distinction prevents development evidence from being silently reinterpreted as confirmation. Table 1 separates exploratory development from locked evaluation. M3CoTBench development was historically inspected and is never presented as confirmatory. The PathVQA cohort excludes images close to earlier project outputs and contains one question per exact-image group. The 3MDBench policy, case budget, aggregator, endpoint, and success rule were hashed before eligible outputs were complete. DDI was likewise prepared and frozen before inference; restricted images and record-level paths are not redistributed.

Table 1: Cohorts and inferential roles. Groups are byte-identical images unless stated otherwise.

Cohort	Cases	Groups	Task	Role
M3CoTBench	514	502	medical MCQ	train/calibration plus historically exposed development
PathVQA	256	256	pathology yes/no	locked dataset transfer
SLAKE	64	64	lesion-mask sensitivity	annotation-guided mechanism control
3MDBench	2,996	2,944	34-way diagnosis	prospective locked falsification
DDI	656	656	malignant/benign	prospective external harm-control test

The primary action set contains four nonoverlapping 2×2 quadrants, each presented jointly with the global image. A fixed-crop baseline uses one crop unconditionally. An exploratory query-conditioned proposal chooses among 13 fixed multiscale windows by pinned CLIP question–crop cosine. The evaluation additionally includes global-plus-two-crop states, neutral RGB images, and SLAKE lesion/control occlusions.

Qwen2.5-VL-3B-Instruct and MedGemma-4B-it are frozen at explicit repository revisions. Inference uses BF16, SDPA, greedy decoding, and at most 32 generated tokens. Model-call identity includes image hash, question, task, rendered state, prompt, model revision, runtime, and decoding parameters. SQLite insertion is immutable. Every policy reuses the exact same global row.

The primary multinomial logistic router predicts help, unchanged, and harm for every action. Features include action identity, task, dimensions, question n-grams and length, global response length, and generated-token confidence statistics. It cannot access crop pixels or responses. Calibration groups set acquisition thresholds. Baselines comprise global-only, fixed crops, confidence and entropy gates, direct signed-gain regression, help-only prediction, matched-count random acquisition, and the restricted action oracle. A pinned AdaptVision-7B adaptation and Qwen2.5-VL-7B resolution controls provide external context.

After the locked PathVQA result, a separate 3MDBench policy is frozen: acquire four quadrant-conditioned responses for the 25% lowest-confidence cases, aggregate all five responses by plurality, and retain global on ties. After that result, a source-only conservative rule is selected without DDI outcomes: replace global only if all four crop-conditioned answers unanimously agree on a different label and their mean confidence exceeds global confidence by at least 0.10. This DDI rule acquires all views and tests corruption control, not efficiency.

Reported metrics include micro-accuracy, macro-F1 where applicable, validity, acquisition and replacement rates, help/harm counts, sequence tokens, and allocated GPU seconds. Paired differences use 10,000 image-group bootstrap resamples for locked analyses (Efron & Tibshirani, 1993). Discordant correctness pairs receive exact McNemar tests. A predeclared family of 32 matched-count router–confidence comparisons uses Holm correction. DDI additionally reports sensitivity, specificity, balanced accuracy, and released skin-tone groups. Image groups do not establish patient independence.

5 Results

5.1 Development gain and locked reversal

Table 2 summarizes the central comparisons. On M3 development, Qwen global-only reaches 34.55%; its router acquires 7.27% and changes no outcome. MedGemma has larger action headroom: its router improves 24.55% to 30.91%, correcting seven answers without corruption ($+6.36$ pp, CI $[+2.59, +11.21]$). This is exploratory because development was historically exposed.

On locked PathVQA, Qwen makes no purchases at its frozen cost threshold. MedGemma instead acquires 91.02% and loses 6.64 pp, correcting five global errors while corrupting 22 correct answers. The four-action oracle reaches 75.39%, 5.08 pp above global, so the failure is not a complete absence of useful views. It is a failure to rank corruption risk under shift.

Table 2: Frozen policy outcomes. Acquisition is the fraction receiving additional views; DDI acquires all views but replaces no answer.

Model	Cohort	Global	Policy	Acquire	Help/Harm	Δ accuracy (95% CI)
Qwen	M3 dev.	34.55	34.55	7.27	0/0	$0.00\ [0.00, 0.00]$
MedGemma	M3 dev.	24.55	30.91	42.73	7/0	$+6.36\ [+2.59, +11.21]$
Qwen	PathVQA	61.72	61.72	0.00	0/0	$0.00\ [0.00, 0.00]$
MedGemma	PathVQA	70.31	63.67	91.02	5/22	$-6.64\ [-10.55, -2.73]$
Qwen	3MDBench	40.92	39.05	25.00	16/72	$-1.87\ [-2.50, -1.27]$
Qwen	DDI	73.93	73.93	100.0	0/0	$0.00\ [0.00, 0.00]$

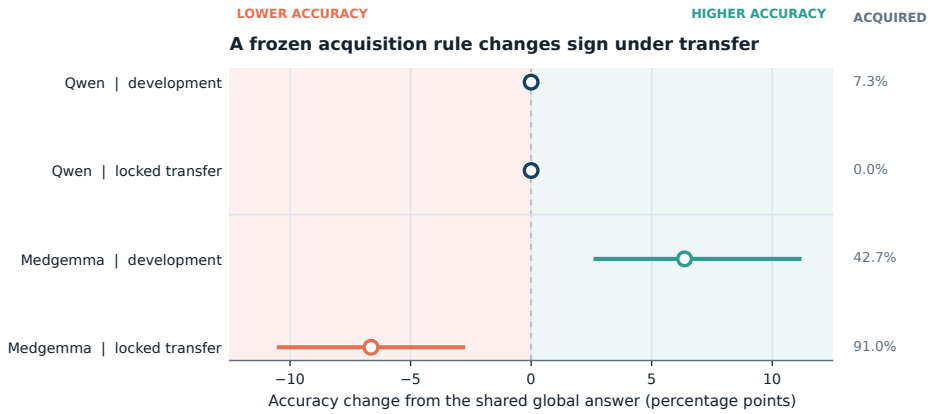


Figure 2: Accuracy change from the shared global response for the frozen cost-sensitive router. M3 development is exploratory; PathVQA is locked.

Figure 2 shows the paired confidence intervals. The sign reversal is model-specific: the same feature design is inert for Qwen and beneficial then harmful for MedGemma. Cross-model transfer is also unstable. Qwen's router purchases nothing on MedGemma, whereas MedGemma's score purchases almost every Qwen case and reduces accuracy by 1.82 pp.

MedGemma's action-level help AUC is 0.685 on development and 0.704 on PathVQA, while harm AUC collapses from 0.624 to 0.255. This asymmetry explains why a score can continue to find some correctable cases but rank harmful actions in the wrong direction. The base rates also reverse (Fig. 3): MedGemma candidate actions are 8.36% helpful and 4.50% harmful on M3 training, but 2.83% helpful and 7.91% harmful on locked PathVQA. The probability of encountering at least one harmful action rises from 7.26% to 12.89%.

At exactly matched acquisition counts, no router–confidence comparison survives Holm correction over the declared 32-comparison family. The MedGemma cost-sensitive router exceeds confidence gating by 4.67 pp on development before correction, but the adjusted p-value is 1.0. Selective prediction similarly yields no threshold with nonzero certified coverage at target risks 0.1–0.4. Accordingly, the evidence supports neither superiority over confidence nor a risk guarantee.

5.2 Prospective stress tests

The 3MDBench rule is frozen before eligible model output. Global Qwen answers 1,226 of 2,996 cases correctly (40.92%). Acquiring four views for the 749 lowest-confidence cases and applying plurality answers 1,170 correctly (39.05%): -1.87 pp with CI $[-2.50, -1.27]$. It helps 16 and harms 72 (exact McNemar $p = 1.19 \times 10^{-9}$) and trails a deterministic random exact-budget policy by 0.90 pp (CI $[-1.56, -0.27]$).

The action oracle reaches 44.46%, leaving 3.54 pp of headroom, yet within selected cases plurality falls from 26.30% global accuracy to 18.83%. Low confidence has AUC 0.571 for helpful flips and 0.638 for harmful flips. It identifies fragility more strongly than recoverability.

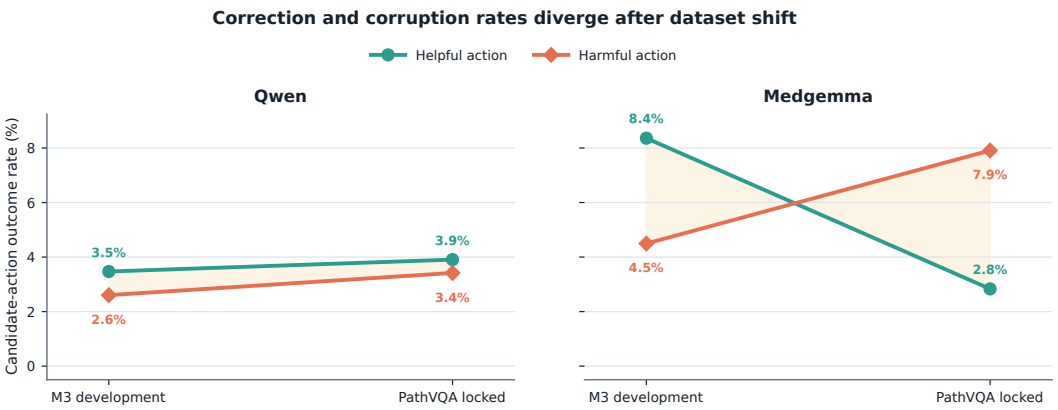


Figure 3: Candidate-action outcome prevalence. MedGemma exhibits corruption-risk reversal: harmful actions become more common than helpful actions under PathVQA shift.

Table 3: Post-hoc aggregation on all 2,996 3MDBench cases. These diagnostics cannot revise the frozen primary result.

Aggregator	Δ accuracy	Help	Harm
Five-view plurality	-3.97 pp	45	164
Crop support ≥ 2	-4.21 pp	46	172
Unanimous crops	-2.40 pp	35	107
Confidence-weighted vote	-4.27 pp	46	174
Maximum-confidence state	-2.30 pp	42	111

Every post-hoc aggregation alternative remains harmful (Table 3). Even unanimous crop voting loses 2.40 pp when always applied. Averaging all 24 crop orders shows that the first plurality view has zero mean value, the second loses 3.38 pp, the third adds 0.01 pp, and the fourth loses 0.60 pp (Fig. 4). These nonmonotone, often negative increments contradict the operational assumption that more inspected regions should steadily improve the answer.

The source-only DDI gate requires unanimous crop predictions and a confidence margin of 0.10. It makes zero replacements, so policy and global accuracy are identically 73.93%. This satisfies sample-level corruption control but fails the predeclared superiority rule and spends five calls per case.

More importantly, aggregate accuracy is misleading. Qwen predicts malignant for 0.30% of cases, producing 0.58% sensitivity, 99.79% specificity, and 50.19% balanced accuracy. Naive plurality reaches 50.58% balanced accuracy; even the action oracle reaches only 52.92%. Across released skin-tone groups, sensitivity is 2.04%, 0%, and 0% (Table 4). Equation 7 localizes this failure: action headroom is too small to support a useful router. The result is neither evidence of clinical competence nor a fairness comparison.

Table 4: Class-conditional DDI performance by released skin-tone group. Accuracy largely tracks benign specificity, whereas malignant sensitivity remains at or near zero. These descriptive groups have uncontrolled prevalence and sampling and therefore do not support a fairness comparison.

Skin-tone group	n	Accuracy	Malignant sensitivity	Benign specificity	Balanced accuracy
12	208	76.92%	2.04%	100.00%	51.02%
34	241	68.88%	0.00%	99.40%	49.70%
56	207	76.81%	0.00%	100.00%	50.00%

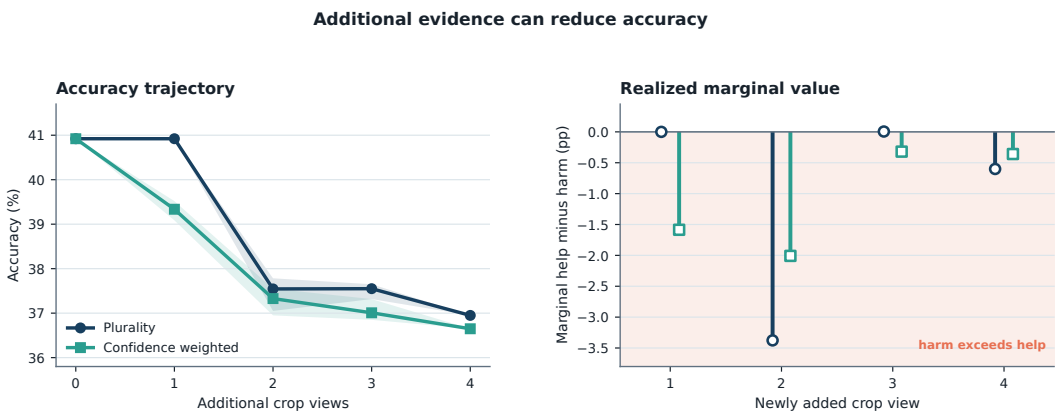


Figure 4: Accuracy and realized marginal value as crop views are added on 3MDBench, averaged over all 24 crop orders. Additional views do not provide monotone gains.

5.3 Visual-use and grounding controls

A routing failure could arise because the VLM ignores images altogether. Controls argue against that simple explanation for MedGemma. Replacing the PathVQA image with constant RGB reduces accuracy by 19.53 pp (CI $[+10.93, +28.52]$). On 64 SLAKE image groups, agreement with the original label is 90.63% for the original and equal-area control occlusion, but 82.81% after lesion-mask occlusion; the control-minus-lesion contrast is $+7.81$ pp (CI $[+1.56, +14.06]$). These interventions show sensitivity to visual and annotated lesion content, not causal grounding or clinical validity. The router fails despite measurable image use. Figure 6 in the appendix shows representative intervention triplets.

The query-conditioned CLIP proposal also fails to solve transfer. For MedGemma development its binary-action oracle is 35.45% and the gate reaches 30.00% at 26.36% acquisition, but matched-count confidence ties it. On PathVQA, always using the proposal reduces accuracy by 5.08 pp and the frozen proposal gate reduces it by 3.91 pp. Better localization does not remove selection risk when action utility shifts.

6 Discussion

The experiments change the interpretation of adaptive visual computation. An accuracy–cost curve is necessary but cannot distinguish a policy that rarely acts from one that finds useful actions, nor a weak selector from an aggregator that discards a correct crop. The paired design separates these cases. Action-oracle headroom measures what the candidate set makes recoverable; help and harm record changes to the shared global answer; selection and aggregation regret locate the lost headroom. This decomposition is the methodological contribution.

The transfer results suggest that correction and corruption should not be treated as opposite labels governed by a single notion of uncertainty. On MedGemma, the help AUC remains similar from development to PathVQA, yet the harm AUC reverses and the base rate of harmful actions rises. Thus a score can retain some ability to find correctable cases while becoming anti-informative about which action is safe. The term *corruption-risk reversal* denotes this observed configuration. It is related to negative transfer (Wang et al., 2019), but is more specific: the object that changes is the conditional sign of an intervention relative to an already generated answer. The evidence spans two model families and several tasks, but it does not imply that every adaptive cropper will exhibit the same reversal, nor that the reported frequencies estimate a clinical population.

The prospective results expose two complementary limits. On 3MDBench, low confidence enriches for changes but those changes are more often harmful than helpful, and plurality cannot remove the error because crop-conditioned responses are correlated. On DDI, unanimity prevents new errors only by declining every replacement. These cases show why safety must be indexed by both risk and coverage. Zero replacement gives a trivial corruption guarantee; it does not reduce compute,

improve prediction, or repair a base model whose candidate actions contain little malignant sensitivity. Aggregate accuracy is especially misleading under class imbalance, a concern that persists when medical imaging models move across populations and sites (Yang et al., 2024).

A more defensible router would estimate action-specific harm under target support rather than treat current-answer confidence as acquisition value. A learned proposal could be combined with a lower confidence bound on signed utility and a support test over pre-acquisition variables. Poor support should lead to retaining or abstaining. Selection and aggregation also require separate training and audit: positive value for one view does not justify voting across correlated decodes. Proposition 2 identifies the need to bound continuation-value error, but this study supplies no distribution-free bound. Conformal control would additionally require a monotone loss and a calibration population representative of deployment.

Several limitations define the claim boundary. M3 development was historically exposed and is hypothesis-generating. Exact-image grouping blocks literal duplicates but cannot establish patient or source independence. Quadrants and a fixed CLIP proposal restrict the action family; the AdaptVision comparison couples checkpoint, resolution, fine-tuning, and generation budget. Generated-token confidence is not an answer probability, and MedGemma's multi-image use is experimental. PathVQA changes domain and answer format; 3MDBench uses the dialogue-free Image + All Complaints condition; DDI lacks patient identifiers. Skin-tone strata are confounded by label prevalence and support no fairness claim. SLAKE occlusion tests sensitivity to annotated content, not causal localization. The experiments establish neither expert-level performance, clinical benefit, deployment safety, nor universal acquisition harm.

7 Conclusion

For an autoregressive medical VLM, a new view intervenes on the predictor as well as observing the image. Its value need not be monotone: a crop may expose a missing cue, remove useful context, or create correlated errors that consensus reinforces. Paired counterfactuals make these outcomes visible and assign failure to the action set, selector, or aggregator rather than one accuracy–cost point.

Across the evaluated transfers and prospective tests, the reliability of acquisition is less stable than its capacity to change an answer. One observed manifestation is termed corruption-risk reversal, with the claim restricted to the studied policies, models, and datasets. The broader lesson is an evaluation requirement: adaptive visual systems should demonstrate recoverable headroom, estimate help and harm separately under target support, preserve useful actions during aggregation, and attain nontrivial coverage. Additional computation is beneficial only when the system can show which changes it should trust.

Acknowledgments

Omitted for double-blind review.

Ethics Statement

This study uses released research datasets and offers no clinical recommendation. Restricted DDI images and record-level metadata are not redistributed. The results establish neither clinical safety, expert equivalence, patient independence, nor fairness; dataset licenses govern all local artifacts.

Reproducibility Statement

The repository provides frozen protocols, hashes, aggregate outputs, analysis code, and 26 passing tests. Restricted DDI data and response caches remain excluded under the dataset agreement.

AI Use Statement

Generative AI assisted coding, debugging, editing, and literature discovery. It did not replace experimental design or scientific judgment; all results were recomputed from saved artifacts, and the authors take responsibility for the final work.

REFERENCES

A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

S. Bai et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

M. Carvalho, H. Dias, and B. Martins. CropVLM: Learning to zoom for fine-grained vision–language perception. *arXiv preprint arXiv:2511.19820*, 2025.

Y. S. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Y. S. Chow, H. Robbins, and D. Siegmund. *Great Expectations: The Theory of Optimal Stopping*. Houghton Mifflin, 1971.

R. Daneshjou et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances*, 8(32):eabq6147, 2022.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1993.

Y. Geifman and R. El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, pp. 2151–2159, 2019.

T. Guan, F. Liu, X. Wu, R. Xian, Z. Li, X. Liu, X. Wang, L. Chen, F. Huang, Y. Yacoob, D. Manocha, and T. Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision–language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

O. B. Guney, K. S. Saichandran, K. Elzokm, Z. Zhang, and V. B. Kolachalama. Active feature acquisition via explainability-driven ranking. In *International Conference on Machine Learning*, pp. 20748–20765, 2025.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017.

P. Hager, F. Jungmann, R. Holland, K. Bhagat, I. Hubrecht, M. Knauer, J. Vielhauer, M. R. Makowski, R. Braren, G. Kaissis, and D. Rueckert. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30:2613–2622, 2024. doi: 10.1038/s41591-024-03097-1.

X. He, Z. Cai, W. Wei, Y. Zhang, L. Mou, E. Xing, and P. Xie. Towards visual question answering on pathology images. In *ACL-IJCNLP*, pp. 708–718, 2021.

J. Jiang et al. M3CoTBench: Benchmark chain-of-thought of MLLMs in medical image understanding. In *International Conference on Learning Representations*, 2026.

Y. Li, Y. Du, K. Zhou, J. Wang, X. Zhao, and J.-R. Wen. Evaluating object hallucination in large vision–language models. In *Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023.

Z. Lin, Y. Liu, Y. Yang, L. Tao, and D. Ye. AdaptVision: Efficient vision–language models via adaptive visual acquisition. *arXiv preprint arXiv:2512.03794*, 2025.

B. Liu, L. M. Zhan, L. Xu, L. Ma, Y. Yang, and X. M. Wu. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *IEEE International Symposium on Biomedical Imaging*, pp. 1650–1654, 2021.

C. Ma, S. Tschiatschek, K. Palla, J. M. Hernandez-Lobato, S. Nowozin, and C. Zhang. EDDI: Efficient dynamic discovery of high-value information with partial VAE. In *International Conference on Machine Learning*, pp. 4234–4243, 2019.

M. S. Ryoo, A. J. Piergiovanni, A. Arnab, M. Dehghani, and A. Angelova. TokenLearner: Adaptive space–time tokenization for videos. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12786–12797, 2021.

A. Sellergren et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

I. Sviridov, A. Miftakhova, A. Tereshchenko, G. Zubkova, P. Blinov, and A. Savchenko. 3MDBench: Medical multimodal multi-agent dialogue benchmark. *arXiv preprint arXiv:2504.13861*, 2025.

M. Valancius, M. Lennon, and J. Oliva. Acquisition conditioned oracle for nongreedy active feature acquisition. In *International Conference on Machine Learning*, pp. 48957–48975, 2024.

Z. Wang, Z. Dai, B. Póczos, and J. Carbonell. Characterizing and avoiding negative transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11293–11302, 2019.

Y. Yang, H. Zhang, J. W. Gichoya, D. Katabi, M. Ghassemi, et al. The limits of fair medical imaging AI in real-world generalization. *Nature Medicine*, 30:2838–2848, 2024. doi: 10.1038/s41591-024-03113-4.

K. Zhang, R. Zhou, E. Adhikarla, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 30:3129–3141, 2024. doi: 10.1038/s41591-024-03185-2.

A Experimental and Reproducibility Detail

A.1 Immutable call identity and pairing

Each response key hashes the rendered evidence specification, image content, prompt, task, model identifier and revision, maximum generated tokens, deterministic decoding configuration, runtime versions, and experiment version. Responses are inserted once. Analysis reconstructs every policy by key and refuses incomplete state coverage. This prevents a policy from silently benefiting from a second global decode or a modified prompt.

The M3 parser accepts answer letters A–Z and scores exact option sets; this corrects an earlier A–F restriction. Binary tasks accept only normalized yes/no prefixes. Invalid outputs remain errors. Open-ended historical analyses use token F1 and are not pooled with MCQ accuracy.

A.2 Compute accounting

The primary M3/PathVQA matrix contains 10,780 paired responses. The complete Qwen 3MD-Bench run contains 14,980 responses (five states per case). DDI adds 3,280 responses. On 3MD-Bench, the frozen policy averages 1,215.8 sequence tokens and 1.070 allocated GPU seconds per case. The all-view DDI protocol averages 1,901 sequence tokens and 1.83 allocated GPU seconds per case. These values divide batch execution time among requests and do not estimate online latency or provider billing.

A.3 Protocol chronology and leakage controls

Table 5 records the inferential role of each phase. "Locked" means that eligible target outputs were unavailable when the policy and thresholds were fixed. M3CoTBench development is retained as exploratory evidence because it was historically exposed. The two prospective phases use distinct policies that were committed before eligible generation began.

Table 5: Protocol chronology. Target labels and post-acquisition responses are never router inputs.

Phase	Policy source	Locked before target	Eligible output seen before freeze	Inferential role
M3 development	Same split	No	Yes	Exploratory construction
PathVQA transfer	M3 only	Yes	No	Locked transfer
3MDBench	Prespecified confidence rule	Yes	No	Prospective stress test
DDI	Prespecified unanimity gate	Yes	No	Prospective safety and coverage test

B Exact Policy Definitions

B.1 Primary signed-utility router

For each action, multinomial logistic models estimate the probabilities of helpful correction, unchanged outcome, and harmful corruption. Inputs are available before acquisition: action identity, task, image dimensions, question length and word n-grams, global-response length, and generated-token confidence summaries. The model cannot inspect crop pixels or crop-conditioned responses. The router selects the action maximizing $\widehat{P}(\mathrm{help}) - \widehat{P}(\mathrm{harm}) - \lambda_a$ and acquires only when the estimated utility passes the development threshold. Every transfer result uses the original feature map, coefficients, thresholds, and action order.

B.2 Baselines and diagnostic oracles

Global-only returns the shared global decode. Fixed-crop always acquires a named action. Confidence and entropy gates acquire from scalar global uncertainty. Matched-count random policies acquire the same number of cases as the learned policy and isolate the value of case selection. Direct signed-gain regression predicts G_a without separating help from harm; the help-only model ignores corruption. The restricted action oracle selects the best answer available among the logged global and candidate actions. It is diagnostic headroom, never a deployable policy.

B.3 Prospective policies

On 3MDBench, the frozen policy acquires four quadrant views for the lowest-confidence 25% of global responses and applies plurality across the global and four acquired answers, retaining the global answer on ties. On DDI, four fixed views are acquired and replacement occurs only when all four views agree on a label different from the global label and their mean confidence exceeds global confidence by at least 0.10. The latter gate made zero replacements, which is interpreted jointly with its zero operational coverage.

C Additional Results

C.1 Oracle headroom and proposal actions

Qwen's four-action oracle reaches 38.18% on M3 development and 67.19% on PathVQA, compared with global accuracies 34.55% and 61.72%. MedGemma's corresponding oracle accuracies are 34.55% and 75.39%, compared with 24.55% and 70.31%. The CLIP proposal oracle reaches 35.45% for MedGemma development and 74.22% on transfer. The proposal improves the development action set slightly but does not prevent transfer harm.

C.2 External adaptive system and resolution control

The AdaptVision-7B adaptation obtains 35.45% accuracy on 110 M3 development questions and requests a crop on 3.64%. Vanilla Qwen2.5-VL-7B obtains 37.27% at full resolution and 30.91% after factor-two downsampling. AdaptVision is -1.82 pp relative to full resolution (CI $[-9.35, +5.61]$) and $+4.55$ pp relative to downsampled input (CI $[-1.89, +11.77]$). The experiment is consistent with recovery of some compression loss but does not isolate a routing advantage.

C.3 Headroom, realized gain, and regret

Table 6 reports a common decomposition. Headroom is oracle minus global accuracy, realized gain is policy minus global accuracy, and residual regret is oracle minus policy accuracy. Negative realized gain means that routing loses accuracy despite the presence of recoverable actions.

Table 6: Action availability and realized routing performance in percentage points.

Model	Dataset	Global	Oracle	Headroom	Gain	Regret
Qwen	M3 dev	34.55	38.18	3.63	0.00	3.63
Qwen	PathVQA	61.72	67.19	5.47	0.00	5.47
MedGemma	M3 dev	24.55	34.55	10.00	+6.36	3.64
MedGemma	PathVQA	70.31	75.39	5.08	−6.64	11.72
Qwen	3MDBench	40.92	44.46	3.54	−1.87	5.41

C.4 Per-diagnosis heterogeneity

Figure 5 displays the descriptive class-level recall change produced by the frozen 3MDBench policy. Gains and losses coexist, while the overall effect is negative. The bars are not multiplicity-adjusted, class sizes vary, and the dataset does not establish patient independence; the figure therefore localizes heterogeneity rather than supporting diagnosis-specific claims.

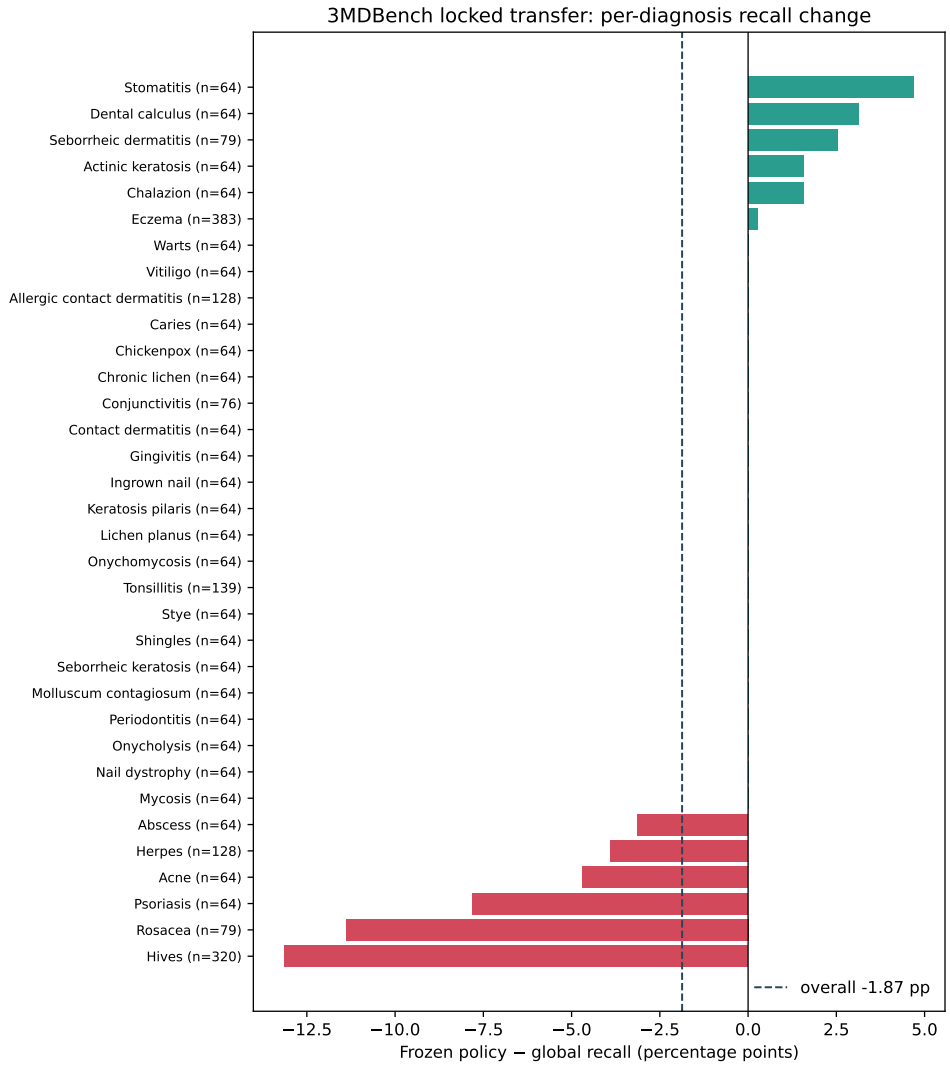


Figure 5: Per-diagnosis recall change under the frozen 3MDBench policy. The dashed line marks the overall -1.87 pp accuracy change.

D Statistical Inference and Claim Audit

All reported confidence intervals use 10,000 image-group bootstrap replicates unless stated otherwise. Paired binary accuracy comparisons use exact McNemar tests. The matched-count family applies Holm correction across 32 comparisons. Image grouping is the strongest unit supported consistently by the released data; it prevents literal image duplication from creating pseudo-replication but cannot substitute for unavailable patient or institution identifiers.

Table 7: Claim audit linking each conclusion to its permitted interpretation.

Finding	Status	Direct evidence	Claim boundary
MedGemma development gain	Exploratory	Paired M3 counterfactuals	Hypothesis generation only
PathVQA reversal	Locked transfer	Frozen M3 router on target data	Studied model, actions, and shift
3MDBench plurality harm	Prospective	2,996 eligible cases	Prespecified policy and input condition
DDI zero-corruption result	Prospective	656 cases, zero replacements	Safety only at zero replacement coverage
Grounding sensitivity	Mechanistic control	Occlusion and constant-image interventions	Sensitivity, not causal localization

E VISUAL-USE AND GROUNDING CONTROLS

Figure 6 illustrates the SLAKE occlusion protocol behind the sensitivity result reported in Section 5.3.

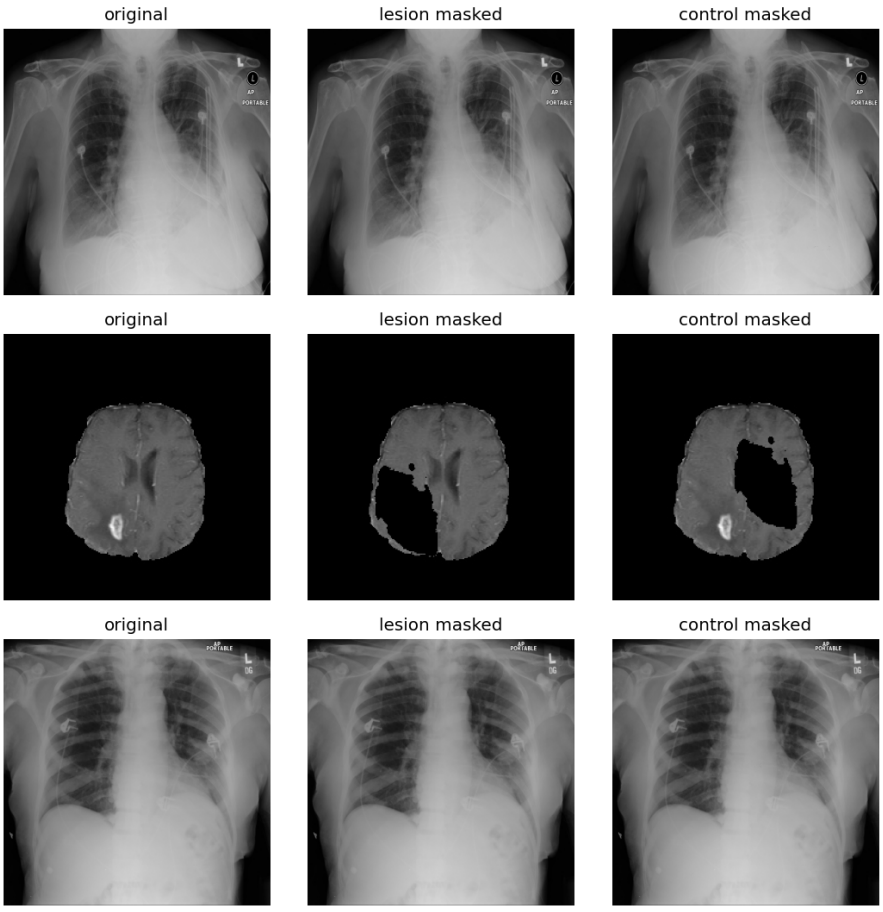


Figure 6: Representative SLAKE intervention triplets. Each row shows the original image, lesion-masked view, and equal-area control-masked view. The montage illustrates construction; the 64-group paired contrast supports the reported sensitivity result.

F REPRODUCIBILITY ARTIFACT MAP

Table 8: Audit trail retained by the released repository.

Artifact	Role	Integrity check
Locked manifests	Exact sample identity and eligibility	Content hashes and fixed counts
SQLite response caches	Immutable model outputs	Compound call identity; insert once
Policy freeze records	Features, coefficients, thresholds, timing	Stored before eligible target generation
Analysis JSON	Metrics, intervals, and decompositions	Recomputed from cache and manifest joins
Compute ledger	Calls, tokens, and allocated GPU seconds	Counts reconciled with complete state grids
Validation suite	Parser, pairing, and policy invariants	26 passing automated tests

Restricted DDI images, record-level metadata, and provider response caches are not redistributed. Their exclusion is a data-governance constraint rather than missing experimental state: public aggregate outputs, manifests, hashes, code, and validation checks preserve the auditable inferential path without releasing restricted records.