SpineFairBench: A Counterfactual Benchmark for Auditing Demographic Sensitivity in Spinal Radiology VLM Reports

Ahmed Taha*
Department of Computer Science
Columbia University
Johns Hopkins University
at4058@columbia.edu

Abdelrahman Taeha
Department of Computer Science
Georgia Tech
ataeha3@gatech.edu

Muzzammil Ahmadzada
School of Medicine
Stanford University
Muzz@Stanford.edu

Abstract

Radiology vision–language models may change clinically meaningful report content when the same pathology is presented under different patient demographics, but observational subgroup audits cannot isolate this effect. We present SpineFairBench, a paired counterfactual benchmark for auditing demographic sensitivity in spinal radiology report generation. SpineFairBench varies apparent age and sex in source/counterfactual spinal radiographs under a target-pathology-preservation criterion. It assesses a frozen nine-model VLM panel under a locked report-generation prompt with two pre-registered primary endpoints: recommendation change rate and diagnostic-label consistency. Retained outputs demonstrate measurable recommendation drift in all nine models. In most retained models, management recommendations are less stable than diagnostic-label overlap under the same demographic edit. A pre-registered findings-first mitigation analysis with a binding interpretation rule produced a negative result on the eligible subset, supporting a predominantly perceptual rather than interpretive locus for the two models on which it could be evaluated. In blinded validation by three board-certified radiologists, clinical plausibility and target-pathology preservation were supported for 443/450 pairs (98.44%) under a 2-of-3 majority rule. Reviewers selected "Cannot tell" for 96.8% of edit-detectability responses.

1 Introduction

Vision-language models (VLMs) are increasingly employed for radiology report generation including spinal imaging. A difficult question to answer in the evaluation of VLMs for clinical use is whether their reports of a patient's spine depend on who the patient appears to be. Recent assessments of VLMs for medical imaging have used correlational analyses of imaging cohorts defined by factors such as age and sex (Yang et al., 2025). These patient cohorts also differ in other variables such as pathology burden, view mix, source institution, and acquisition quality. The gaps identified in these analyses are informative of potential disparities, but cannot directly establish that those disparities are due to demographic presentation itself. VLMs are not evaluated under paired counterfactual conditions in which target pathology is preserved and demographic presentation in the image is varied. This is a limitation of the study design, not the data: more observational data cannot separate these confounds.

Counterfactual evaluation differs in what is compared: the same case under different demographic presentations, rather than separate patient cohorts. We take the same source image and edit it to

*Code: https://github.com/ahmedtaha100/SpineFairBench.
Data: https://huggingface.co/datasets/ahmedtaha100/spinefairbench-artifacts.

Preprint.

change apparent age and sex while preserving target pathology, then compare reports generated for the source and edited image directly.

Counterfactual image generation has begun to appear in medical imaging for demographic editing and downstream discriminative modeling (Pombo et al., 2023; Yeganeh et al., 2025), but has not been systematically applied to the evaluation of VLMs generating free-text reports of spinal radiographs, where clinically relevant shifts may appear in recommendations, severity language, confidence, refusals, and hallucinated findings even when diagnostic labels remain stable.

To determine whether the reports generated by VLMs change under controlled demographic counterfactual edits, we introduce SpineFairBench. For each source radiograph, up to four target demographic conditions are edited in, and only source/counterfactual pairs that pass automated quality control are included in the benchmark. Under a locked reporting prompt, a frozen panel of VLMs is evaluated on each pair. The headline claims are based on a verified core taken from two open access spine datasets with trusted age and sex metadata (Pham et al., 2021; Nguyen et al., 2021; Klinwichit et al., 2023; Faculty of Informatics, Burapha University).

SpineFairBench is explicitly limited to apparent age and sex in spine X-rays; race, ethnicity, other modalities, and downstream clinical outcomes are beyond the scope of the current claim. The operative notion of counterfactual validity is *target pathology preservation under controlled demographic editing*, not perfect isolation of every demographically correlated radiographic feature: age and sex in spinal radiographs have genuine radiographic correlates that no image editing pipeline can fully decouple from pathology-adjacent anatomy. Blinded radiologist validation on a stratified post-QC subset supports pathology preservation and clinical plausibility (Section 5).

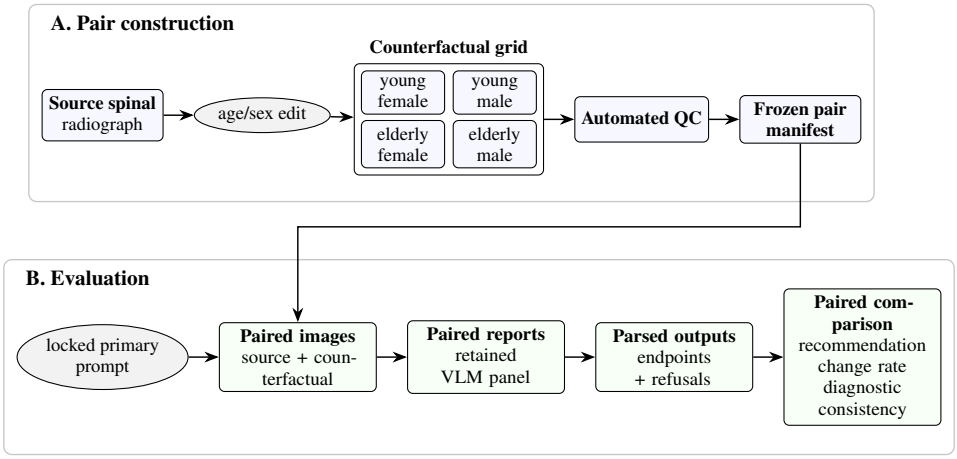

Figure 1: SpineFairBench constructs source/counterfactual spinal-radiograph pairs by editing apparent age and sex, applying automated QC, and freezing the retained pair manifest. Retained VLMs generate reports for source and counterfactual images under a locked prompt, reports are parsed into endpoints and refusals, and source vs. counterfactual reports are compared on the two primary endpoints (recommendation change rate, diagnostic-label consistency).

Contributions. This paper makes the following contributions.

1. **A counterfactual spinal-radiograph dataset and benchmark.** We release counterfactual spinal radiographs paired to source-image identifiers, with frozen pair manifests and automated QC metadata. Source radiographs remain under VinDr-SpineXR and BUU-LSPINE access terms.
2. **A pre-specified VLM evaluation protocol.** We evaluate a frozen nine-model retained panel under a locked report-generation prompt with two primary endpoints: recommendation change rate and diagnostic-label consistency.
3. **A validity framework for counterfactual report auditing.** We define automated QC, blinded radiologist validation, refusal-aware accounting, and paired statistical analysis with source-clustered bootstrap intervals.

4. **Empirical findings across nine VLMs.** Recommendation change is observed in all nine retained models; in most retained models, management recommendations are less stable than diagnostic-label overlap under the same demographic edit.
5. **A pre-specified mechanism and mitigation analysis.** We test whether report drift has a perceptual or interpretive locus through a findings-first decomposition with a binding interpretation rule; the negative outcome on the eligible subset supports a predominantly perceptual rather than interpretive locus for the two models on which it could be evaluated.

2 Related Work and Positioning

Observational demographic audits in medical imaging. In the medical imaging context, fairness work mainly compares performance across observed subgroups defined by age, sex, race, or ethnicity. Recent audits of expert-level vision–language foundation models report demographic disparities in chest-X-ray diagnosis tasks (Yang et al., 2025). This work establishes demographic disparities in radiology VLMs but cannot, by design, show that demographic presentation itself is the cause.

Counterfactual auditing and controlled demographic editing. Counterfactual fairness shifts the question to whether the same case would receive a different model output under a controlled demographic intervention (Kusner et al., 2017). This idea has been applied to medical-imaging work with generative counterfactuals, including morphologically constrained 3D brain-MRI counterfactual augmentation for downstream discriminative modeling and, more recently, diffusion-based methods for structure-preserving medical-image generation (Pombo et al., 2023; Huang et al., 2026; Yeganeh et al., 2025). Although such approaches promote paired image-level interventions, we are not aware of any evaluation benchmark for demographically driven report instability in free-text spinal-radiology VLM reports.

Fairness benchmark infrastructure and radiology-report evaluation. Existing medical fairness benchmarks are primarily designed for discrete classification. MEDFAIR provides standardized datasets, debiasing algorithms, and fairness metrics for medical-image prediction (Zong et al., 2023), but not generated reports. Radiology-report evaluation has progressed from surface-language overlap to clinical labels, entity graphs, and composite radiologist-aligned metrics (Smit et al., 2020; Jain et al., 2021; Yu et al., 2023). None of these directly measure whether recommendations, diagnostic labels, severity language, confidence, refusals, or hallucinations change under edited demographic counterfactuals.

Positioning of SpineFairBench. SpineFairBench brings together four elements that, to our knowledge, have yet to be jointly utilized: a controlled counterfactual intervention at the level of the spinal-radiograph image; free-text VLM reports as the evaluation surface; pre-specified primary endpoints for recommendation change and diagnostic-label consistency; and refusal-aware common-pairs accounting. The contribution of this paper is evaluative: a benchmark to measure whether spinal-radiology VLM behavior changes when the same pathology is presented under edited demographic counterfactuals.

3 What SpineFairBench Evaluates

Scientific role and scope. SpineFairBench is an evaluation benchmark, not a report-generation system. Each evaluation compares reports generated for the same spinal radiograph before and after a controlled apparent-age or sex edit, with target pathology preserved.

Non-claims and causal validity. SpineFairBench does not claim broad clinical fairness in radiology, patient-care effects, generalization beyond spine X-rays, or a perfectly isolated intervention on apparent age and sex. Age and sex leave genuine radiographic signatures in the spine, including bone density, vertebral body morphology, endplate sclerosis, and kyphotic angle. The benchmark therefore defines counterfactual validity as *target-pathology preservation under the edit*, not isolation of every demographically correlated feature. This section formalizes this causal assumption, while radiologist validation and automated QC test pathology preservation.

Audited output surface and evidence commitments. SpineFairBench audits free-text reports rather than a single classification label because a report can preserve a diagnosis while changing recommendations, severity language, confidence, refusals, or hallucinated findings. The primary endpoints are recommendation change rate and diagnostic-label consistency; severity and confidence language are secondary; hallucination is exploratory and restricted to the VinDr trusted-label subset. Primary claims use paired comparisons under the locked prompt with refusal-aware accounting and pre-specified evidence rules. Paired-report instability under the edit operationalizes demographic sensitivity for free-text report generation; it is not a definition of clinical fairness or downstream harm.

4 Benchmark Construction

4.1 Source datasets

SpineFairBench uses a verified core from two open-access spinal-radiograph datasets with trusted per-case age and sex metadata: VinDr-SpineXR (Pham et al., 2021; Nguyen et al., 2021) and BUU-LSPINE (Klinwichit et al., 2023; Faculty of Informatics, Burapha University). These contribute 3,000 raw verified-core sources.

4.2 Counterfactual generator

The generator uses a Stable Diffusion v1.5 backbone (Rombach et al., 2022) with LoRA adapters (Hu et al., 2022) on U-Net attention blocks (rank 64, alpha 128). The retained checkpoint was trained on 9,024 spine radiographs from VinDr-SpineXR and BUU-LSPINE using a single H200 GPU. Stage 1 optimizes an L1 denoising/reconstruction loss with weight 1.0, PatchGAN adversarial generator loss with weight 0.005, and KL regularization with weight 10^{-7}; the discriminator loss averages real and fake PatchGAN binary-cross-entropy terms. Stage 2 keeps the same generator/discriminator terms and turns on latent cycle-consistency L1 loss with weight 1.0.

The production img2img inference path encodes the source image into latent space using the Stable Diffusion variational autoencoder (VAE), applies 50 denoising steps under guidance scale 5.0, strength 0.15, seed 42, and four locked demographic prompt templates: "Lumbar spine X-ray of a 75-year-old female patient," "Lumbar spine X-ray of a 75-year-old male patient," "Lumbar spine X-ray of a 25-year-old female patient," and "Lumbar spine X-ray of a 25-year-old male patient." TSXR masks (Alshenoudy et al., 2026), an X-ray extension of TotalSegmentator2D (Sabrowsky-Hirsch et al., 2026), anchor vertebral and disc regions by blending the latent back toward the source at mask blend 0.7 per denoising step while allowing demographic edits outside the mask. Generator training and benchmark evaluation draw from the same two source datasets.

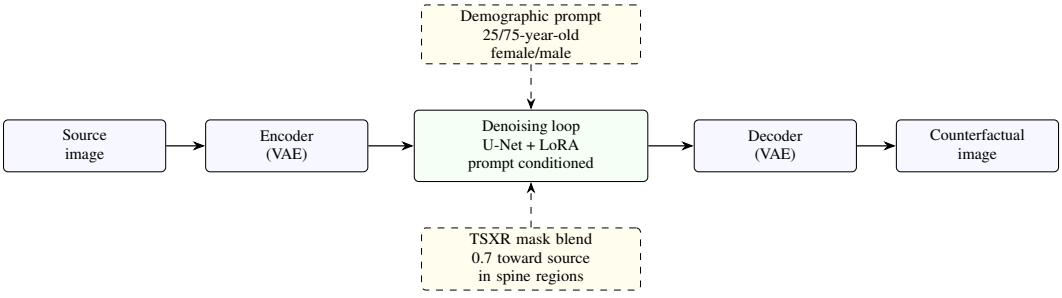

Figure 2: Generator inference path. Source images are encoded, denoised under demographic prompt conditioning with TSXR mask blending preserving pathology-bearing regions, and decoded into counterfactual outputs.

4.3 Pipeline and quality control

Construction proceeds in pipeline order. First, TSXR Rule 1 drops any source with fewer than four detected vertebrae, excluding 13 sources and leaving 2,987 filtered sources. Second, the generator applies four demographic edits per source (young female, young male, elderly female, and elderly

male), yielding approximately 11,948 candidate pairs. Third, automated pair-level QC applies the frozen thresholds SSIM ≥ 0.70, edge preservation ≥ 0.276 with 3×3 dilation, and LPIPS ≤ 0.40; pairs failing any threshold are excluded without alternative-seed regeneration, leaving 11,795 QC-passed pairs spanning 2,950 sources. The 37-source difference from the filtered source count reflects sources with no retained QC-passed counterfactuals. Fourth, a frozen 2,000-source evaluation subset is sampled from this 2,950-source QC-passed pool for compute tractability across the nine-model panel; four edge-preservation drops yield 7,996 evaluable pairs. Fifth, a deterministic 1,000-source common core is drawn from the evaluation subset, yielding up to about 3,998 per-model usable pairs after refusal-aware filtering. Finally, the all-model intersection yields 2,166 cross-model headline pairs.

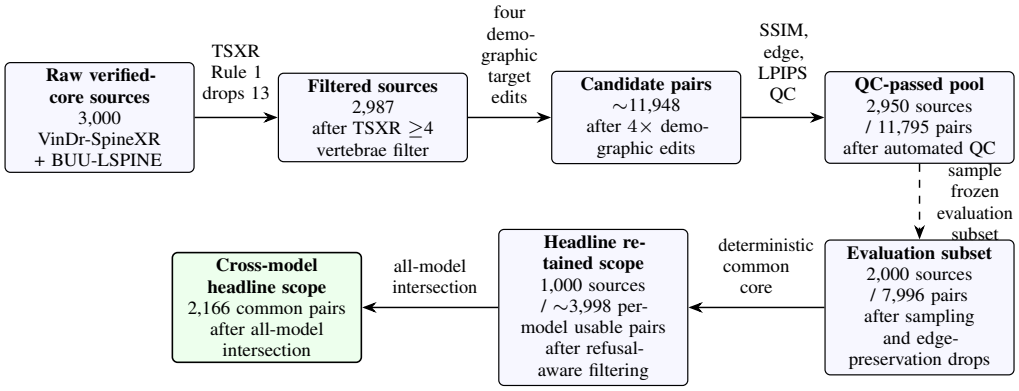

Figure 3: SpineFairBench construction pipeline. Source counts after each filtering, generation, and sampling stage, ending with the cross-model headline scope used in Table 1.

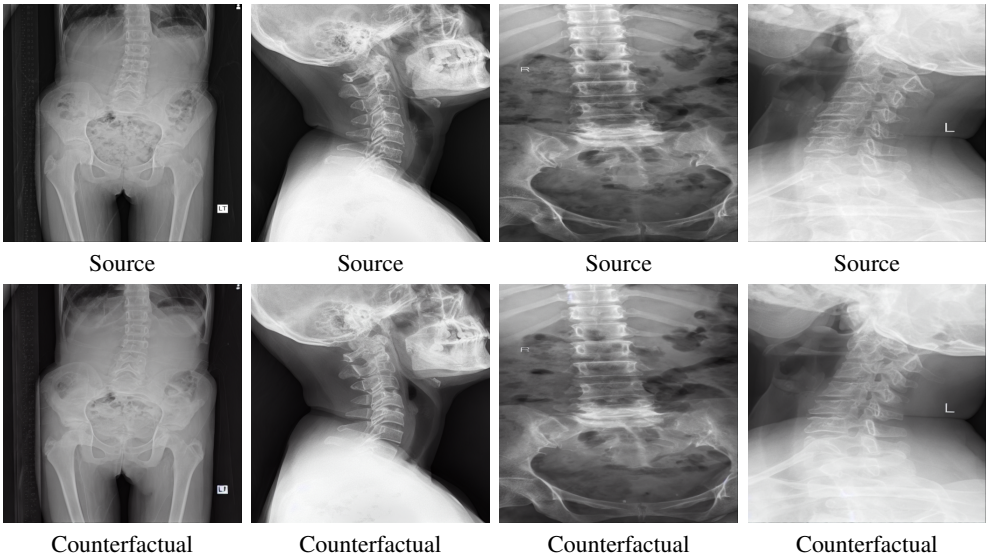

Figure 4: Representative QC-passing source/counterfactual pairs from the release benchmark.

4.4 Common-core composition

The 1,000-source common core is the per-source population on which headline retained analyses are computed. It includes 181 VinDr-SpineXR sources and 819 BUU-LSPINE sources, with all counts in this paragraph reported per source. The sex split is 659 female and 341 male sources, and the age-bucket split is 98 young, 485 middle, and 417 elderly sources. Under the source-demographics manifest pathology labels, 906 sources are no-finding cases and 94 are any-finding

cases; no common-core sources have missing pathology labels. No-finding cases provide the cleanest test of demographic-only drift because no pathology is present to confound the comparison; any-finding cases test preservation under the additional constraint that diagnostic content remain stable.

5 Pair-Construction Validation

Radiologist validation of pair construction. A blinded validation branch evaluated pathology preservation, clinical plausibility, pair quality, and edit detectability on 450 stratified post-QC pairs plus 10 hidden repeats per reviewer. Three board-certified radiologists independently reviewed all pairs. Under the pre-specified 2-of-3 rule requiring both clinical plausibility and pathology preservation, 443/450 pairs (98.44%) passed validation. Across 1,350 detectability responses, reviewers selected "Cannot tell" in 1,307 responses (96.8%). The 450-pair stratified validation sample supports the radiologist-validated subset tier under pre-committed criteria; it is not claimed to be a representative random sample of the 7,996-pair model-evaluation pool.

6 Evaluation Protocol

Evaluation tiers and robustness checks. Per-model headline estimates in Table 2 are reported on each model's refusal-aware analyzable-pair denominator (Table 1) after full-refusal exclusion. Because models refuse different pairs, per-model denominators differ and per-model endpoints are not strictly comparable across models. The 2,166-pair all-model intersection—the subset of pairs on which every retained model produced analyzable output—provides a common denominator for cross-model comparisons at the cost of sample size. The 1,000-source common core is the fixed reference set used for headline summaries and reproducibility. Stricter QC subsets, the radiologist-validated subset, and the all-model intersection serve as robustness checks.

Table 1: Retained model panel and pair accounting. Panel columns report mitigation-attempt eligibility; accounting columns report per-model evaluated pairs, full refusals, analyzable pairs, and the 2,166-pair all-model intersection.

Model	Sub-panel	Counterfactual benchmark	Mitigation attempt	Pairs evaluated	Full refusals	Analyzable
`gpt-5.4`	full-pipeline	✓	✓	3,998	0	3,998
`claude-sonnet-4-6`	full-pipeline	✓	✓	3,998	0	3,998
`claude-opus-4-6`	full-pipeline	✓	✓	3,998	0	3,998
`glm-4.6v`	full-pipeline	✓	✓	3,998	10	3,988
`kimi-k2.5`	full-pipeline	✓	✓	3,998	4	3,994
`gemma-4`	full-pipeline	✓	✓	3,998	1,822	2,176
`llama-4-scout`	baseline-only	✓		3,998	0	3,998
`qwen2.5-vl`	baseline-only	✓		3,998	0	3,998
`radfm`	baseline-only	✓		3,998	0	3,998
All-model intersection	–	–	–	–	–	2,166

Retained-panel rationale. The retained analysis carries nine models (Table 1), frozen on 2026-04-20 and split into full-pipeline and baseline-only sub-panels. `gemini-2.5-pro` and `medgemma` were excluded under dated artifact-readiness and clean-freeze gates before retained analysis: `gemini-2.5-pro` produced usable text on 0 of 10 configured live-client checks, and `medgemma`'s fresh 25-source canary was stopped after 11 of 200 expected calls, with all 11 completed outputs parsed as partial generic-template outputs. These exclusions were determined before retained-analysis launch or final-panel inclusion, not from observed endpoint values. `llama-4-scout` and `qwen2.5-vl` ran the main benchmark cleanly but are baseline-only because their mitigation Stage-1 outputs showed artifact contamination that made them unusable downstream. `radfm` is baseline-only because it produced unreliable outputs on the text-only mitigation stages that the findings-first design requires.

Prompting, parsing, and pair accounting. Every model is given the same fixed prompt, which asks for findings, primary diagnosis, severity, recommended next steps, and confidence. The free-text reply is then parsed into five fields used for evaluation: recommendation, diagnosis, severity, confidence, and hallucination. Refusals are handled explicitly: if a model refuses entirely or returns an API error, that pair is dropped for that model; partial refusals are kept and counted separately.

Endpoints and statistical plan. Two primary endpoints carry the headline claims. Recommendation change rate is the fraction of source/counterfactual pairs where the parsed recommendation changes between the two reports. Diagnostic-label consistency measures how much the diagnosis text overlaps between the two reports, computed as the average Jaccard overlap of their tokenized diagnosis fields. Severity and confidence language are secondary endpoints; hallucination is exploratory and reported only on the VinDr subset where trusted labels are available. All confidence intervals in the main paper are bootstrap intervals with 10,000 resamples, resampling at the source-radiograph level (each draw keeps all four counterfactual pairs from a sampled source together).

7 Main Results

Primary endpoints. Verified retained outputs show recommendation change in all nine models (Table 2). Full-pipeline recommendation change rates range from 0.62 to 0.91; baseline-only models range from 0.29 to 0.57. Diagnostic-label consistency is imperfect in every model. `gemma-4`'s numbers are computed on 2,176 pairs after excluding 1,822 full refusals (46% of pairs); the 280 partial refusals in the remaining set were kept. On the 2,166-pair intersection, gemma-4's values (0.626 / 0.299) are essentially identical to its per-model values, so refusal exclusion does not materially shift its estimates.

Table 2: Retained endpoint summary. Primary endpoints are recommendation change rate and diagnostic-label consistency; severity and confidence are secondary mean absolute paired differences; hallucination is the exploratory VinDr trusted-label disparity. 95% confidence intervals are source-clustered bootstrap intervals; pair accounting is in Table 1.

Model	Sub-panel	Rec. change	Rec. 95% CI	Diag. consistency	Diag. 95% CI	Severity	Confidence	Hallucination
`kimi-k2.5`	full	0.911	[0.902, 0.920]	0.529	[0.519, 0.539]	0.310	1.497	0.089
`claude-sonnet-4-6`	full	0.899	[0.890, 0.909]	0.537	[0.529, 0.545]	0.350	1.804	0.031
`claude-opus-4-6`	full	0.816	[0.802, 0.829]	0.638	[0.629, 0.647]	0.172	1.804	0.012
`glm-4.6v`	full	0.702	[0.686, 0.718]	0.560	[0.551, 0.568]	0.378	0.561	0.174
`gpt-5.4`	full	0.694	[0.678, 0.710]	0.649	[0.640, 0.657]	0.411	1.129	0.011
`gemma-4`	full	0.624	[0.604, 0.644]	0.299	[0.285, 0.312]	1.337	0.925	0.314
`llama-4-scout`	baseline	0.567	[0.548, 0.586]	0.596	[0.585, 0.607]	0.397	1.536	0.448
`radfm`	baseline	0.318	[0.290, 0.346]	0.263	[0.241, 0.286]	0.969	0.447	0.419
`qwen2.5-vl`	baseline	0.293	[0.271, 0.318]	0.545	[0.525, 0.565]	0.053	0.645	0.327

Diagnostic–recommendation stability gap. Across the retained panel, recommendations change more under demographic edits than diagnoses do. For each model, we compute the gap between diagnostic-label consistency and recommendation stability (1 minus the recommendation change rate); the median across models is 0.262. Across full-pipeline models, recommendations change more than diagnoses under the same demographic edit. This matters because recommendations are closer to clinical action than diagnostic labels. The pattern holds in 5 of 6 full-pipeline models and 6 of 9 retained models. In three models, the pattern is reversed – diagnoses change more than recommendations: `gemma-4` (-0.077), `qwen2.5-vl` (-0.162), and `radfm` (-0.419). The pattern holds on the 2,166-pair all-model intersection: 5 of 6 full-pipeline models and 6 of 9 retained models, with median gaps of 0.391 (full-pipeline) and 0.266 (all retained).

Secondary and exploratory endpoints. Beyond the two primary endpoints, we report three additional metrics for context, not as headline claims. Severity and confidence metrics measure how much clinical severity assessment and certainty wording shift under the demographic edit; the largest confidence shifts come from `claude-opus-4-6` and `claude-sonnet-4-6` (Table 2). Hallucination, reported only on the VinDr subset where trusted labels exist, measures how often the model invents findings not actually present; the largest disparities concentrate in the baseline-only models – `llama-4-scout` (0.448), `radfm` (0.419), and `qwen2.5-vl` (0.327).

8 Mitigation Pipeline

The mitigation analysis asks whether demographic drift comes from how the model reads the image or from how it writes the recommendation.

Condition A. The baseline gives the model each source or counterfactual image under the locked primary prompt, which asks for findings, primary diagnosis, severity, recommended next steps, and confidence. The resulting free-text report is parsed into the retained diagnosis and recommendation endpoints.

Condition B. The findings-first path gives the image to the model only in Stage 1, where the prompt asks for anatomical location, finding description, normal/abnormal status, abnormal severity, primary diagnosis, overall severity, and confidence. Stage 2 is text-only: the model receives the same structured findings text, has no image access, and generates severity, recommended next steps, and confidence from those findings.

Condition B′. This extension inserts a deterministic step between Stage 1 and Stage 2 that maps the free-text findings onto a fixed list of standard diagnostic labels. It tests whether removing free-text variation in how findings are described changes the recommendation behavior.

Condition D. This extension rewrites the Stage 2 recommendation draft to remove sentences unsupported by the structured findings. It tests whether stricter output grounding changes the recommendation behavior.

Rule, eligibility, and result. The binding rule, committed on 2026-04-08 in a frozen internal artifact, says Condition B supports an interpretive locus only if recommendation change falls and diagnostic-label consistency drops by no more than 0.05 from Condition A. In other words: if removing the image at the recommendation step makes recommendations more stable without degrading the diagnoses, the drift was coming from how the recommendation is written; if recommendations become less stable or diagnostic consistency drops by more than 0.05, the drift was already in how the image was read. The rule applies only to models whose Stage 1 parse rate is at least 95%, because an unparseable Stage 1 cannot drive a faithful Stage 2. Condition B entered the mitigation design for all six full-pipeline models; `gemma-4` produced only partial artifacts and was excluded before mitigation analysis, while only `gpt-5.4` and `glm-4.6v` passed the parse gate (96.5% each). Sub-threshold failures were almost entirely format violations (claude-sonnet-4-6 77%, claude-opus-4-6 98.5%, kimi-k2.5 100%), meaning the findings-first prompt was hard to elicit—not evidence for either locus. In both eligible models, recommendation change rose rather than fell and diagnostic-label consistency dropped beyond the guardrail: `gpt-5.4` moved from 0.692 to 0.736 in recommendation change and from 0.649 to 0.555 in diagnostic consistency, while `glm-4.6v` moved from 0.708 to 0.861 and from 0.560 to 0.468. Under the binding rule, this supports a predominantly perceptual rather than interpretive locus for these two models only, meaning the drift comes from how the image is read, not from how the recommendation is written.

Table 3: Mitigation status, all full-pipeline models. Top block: models passing the 95% Stage-1 parse gate, evaluated under the binding rule. Bottom block: models not interpreted (Stage 1 below threshold or excluded before analysis). Condition A is recomputed on the Stage-1 parse-gate–eligible subset for matched comparison with Condition B; deltas use unrounded paired values.

Model	Stage-1	Status	Rec. A	Rec. B	Δ rec	Diag. A	Diag. B	Δ diag
`gpt-5.4`	96.5%	binding rule not satisfied	0.692	0.736	+0.043	0.649	0.555	−0.094
`glm-4.6v`	96.5%	binding rule not satisfied	0.708	0.861	+0.154	0.560	0.468	−0.092
Sub-threshold (Stage 1 < 95%): not interpreted under the binding rule.								
`kimi-k2.5`	38.5%	Stage 1 below threshold	0.911	–	–	0.529	–	–
`claude-opus-4-6`	3.0%	Stage 1 below threshold	0.816	–	–	0.638	–	–
`claude-sonnet-4-6`	0.0%	Stage 1 below threshold	0.899	–	–	0.537	–	–
`gemma-4`	–	excluded before analysis (partial-artifact run)	0.624	–	–	0.299	–	–

Conditions A (baseline) and B (findings-first) per Section 8. Binding rule (committed on 2026-04-08 in a frozen internal artifact): satisfied only if $\Delta rec < 0$ and $\Delta diag \geq -0.05$. Both interpreted models fail both halves.

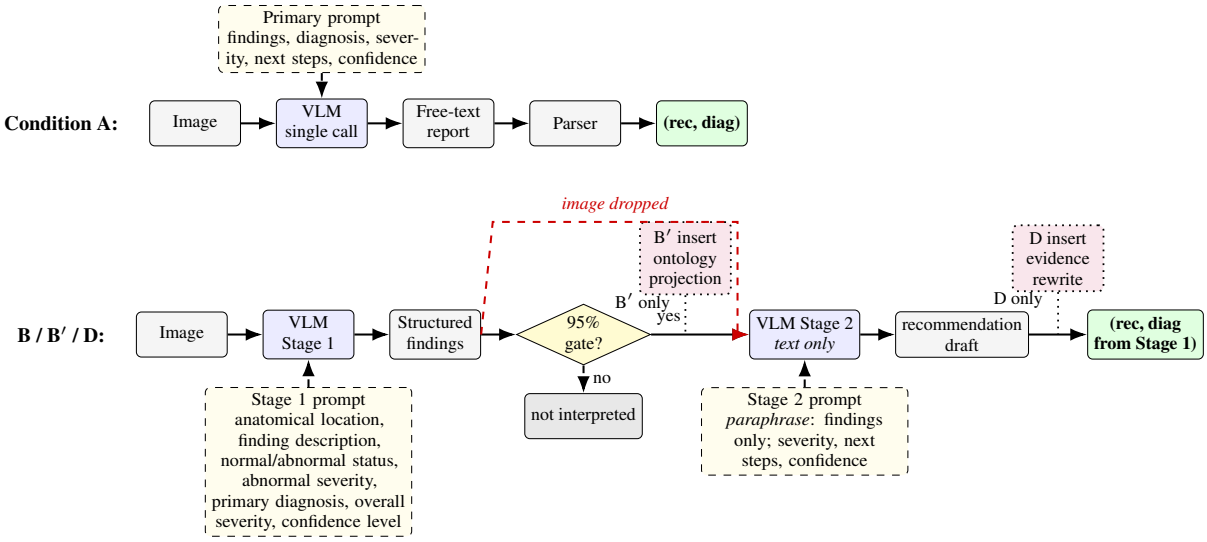

Figure 5: Mitigation flow. Condition A is the baseline single-call report; Conditions B, B′, and D share a Stage 1 → 95% parse gate → text-only Stage 2 path, with B′ and D as optional deterministic inserts.

9 Limitations, Governance, and Reproducibility

Scope and validity. SpineFairBench tests whether spinal-radiology VLM reports change when apparent age or sex is edited while target pathology is preserved. It does not measure clinical fairness, patient outcomes, deployment readiness, race or ethnicity, or behavior outside spine X-rays. Age and sex have real radiographic correlates, so the validity criterion is target-pathology preservation, checked by automated QC and blinded radiologist review (Section 5). Automated QC and blinded radiologist validation reduce, but cannot eliminate, the possibility of synthetic-edit artifacts. Recommendation-change rates reflect paired source/counterfactual instability, not separation from model stochasticity.

Data, labels, and contamination. Hallucination is reported only on the VinDr trusted-label subset; BUU-LSPINE labels are not verified the same way. Race and ethnicity are not reliably recorded in the source datasets and are out of scope. Generator training and benchmark evaluation use overlapping source pools; this is acceptable because the test compares a VLM's reports on the same image before and after editing, not its generalization to unseen images. Provider-served VLMs may also have encountered public spine datasets during pretraining, and exact provider model snapshots and some inference parameters were not under our control.

Governance and reproducibility. SpineFairBench is an audit instrument, not model certification. Synthetic counterfactuals are released for evaluation only and must not be used as authentic clinical images or training data. Reported results trace to frozen manifests, locked prompts and parsers, refusal-aware accounting, and source-clustered bootstrap intervals. Code and reproduction scripts, including SHA-256 checksums and a quickstart that recomputes a representative Table 2 row, are released at `https://github.com/ahmedtaha100/SpineFairBench`. Counterfactual images, manifests, retained model-output summaries, the public radiologist-validation files, and Croissant metadata are released at `https://huggingface.co/datasets/ahmedtaha100/spinefairbench-artifacts`. Source radiographs are not redistributed and remain governed by VinDr-SpineXR and BUU-LSPINE access terms.

10 Conclusion

Every model in the nine-model panel changed its recommendations when the same image was edited to show a different age or sex. In 5 of the 6 full-pipeline models, recommendations changed more than

diagnostic labels did under the same edit, with a panel-wide median gap of 0.262 between diagnostic-label consistency and recommendation stability. This matters because recommendation language is closer to what a clinician would actually do than the diagnostic label alone. The pre-specified mitigation tested whether splitting the task—produce structured findings from the image first, then write the recommendation from those findings as text only—would reduce this drift. On the two models where the structured-findings step parsed reliably enough to evaluate, splitting the task made recommendations less stable, not more, and also reduced diagnostic consistency below the threshold we had committed to in advance. Under our pre-specified rule, this means the demographic effect for these two models comes mostly from how the image is read, not from how the recommendation is written. The release includes counterfactual images, source-image identifiers, model outputs, prompts and parsers, the 443-pair radiologist-validated subset, and the mitigation outputs needed to verify these findings.

References

Ahmed Alshenoudy, Bertram Sabrowsky-Hirsch, Stefan Thumfart, and Michael Giretzlehner. Leveraging synthetic data for whole-body segmentation in x-ray images. In Sharib Ali, David C. Hogg, and Michelle Peckham, editors, *Medical Image Understanding and Analysis*, volume 15916 of *Lecture Notes in Computer Science*, pages 145–158, Cham, 2026. Springer Nature Switzerland. doi: 10.1007/978-3-031-98688-8_11.

Faculty of Informatics, Burapha University. BUU spine dataset. URL `https://services.informatics.buu.ac.th/spine/`.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*, April 2022. URL `https://www.microsoft.com/en-us/research/publication/lora-low-rank-adaptation-of-large-language-models/`.

Fangrui Huang, Alan Wang, Binxu Li, Bailey Trang, Ridvan Yesiloglu, Tianyu Hua, Wei Peng, and Ehsan Adeli. Cycle Diffusion Model for Counterfactual Image Generation. In Islem Rekik, Ehsan Adeli, Sang Hyun Park, and Celia Cintas, editors, *Predictive Intelligence in Medicine*, volume 16164, pages 173–185, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-07904-6. doi: 10.1007/978-3-032-07904-6_16.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q. H. Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. In *Proceedings of the Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Podchara Klinwichit, Watcharaphong Yookwan, Sornsupha Limchareon, Krisana Chinnasarn, Jun-Su Jang, and Athita Onuean. BUU-LSPINE: A Thai Open Lumbar Spine Dataset for Spondylolisthesis Detection. *Applied Sciences*, 13(15):8646, 2023. ISSN 2076-3417. doi: 10.3390/app13158646. URL `https://www.mdpi.com/2076-3417/13/15/8646`.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html`.

Hieu T. Nguyen, Hieu H. Pham, Nghia T. Nguyen, Ha Q. Nguyen, Thang Q. Huynh, Minh Dao, and Van Vu. VinDr-SpineXR: A Deep Learning Framework for Spinal Lesions Detection and Classification from Radiographs. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 291–301, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87240-3. doi: 10.1007/978-3-030-87240-3_28.

Hieu Huy Pham, Hieu Nguyen Trung, and Ha Quy Nguyen. VinDr-SpineXR: A large annotated medical image dataset for spinal lesions detection and classification from radiographs, 2021. doi: 10.13026/q45h-5h59. URL https://physionet.org/content/vindr-spinexr/1.0.0/.

Guilherme Pombo, Robert Gray, M. Jorge Cardoso, Sebastien Ourselin, Geraint Rees, John Ashburner, and Parashkev Nachev. Equitable modelling of brain imaging by counterfactual augmentation with morphologically constrained 3D deep generative models. *Medical Image Analysis*, 84: 102723, February 2023. ISSN 1361-8415. doi: 10.1016/j.media.2022.102723. URL https://www.sciencedirect.com/science/article/pii/S1361841522003516.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Bertram Sabrowsky-Hirsch, Ahmed Alshenoudy, Stefan Thumfart, and Michael Giretzlehner. TotalSegmentator 2D: A Tool for Rapid Anatomical Structure Analysis. In Sharib Ali, David C. Hogg, and Michelle Peckham, editors, *Medical Image Understanding and Analysis*, pages 32–43, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-031-98694-9. doi: 10.1007/978-3-031-98694-9_3.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.117. URL https://aclanthology.org/2020.emnlp-main.117/.

Yuzhe Yang, Yujia Liu, Xin Liu, Avanti Gulhane, Domenico Mastrodicasa, Wei Wu, Edward J. Wang, Dushyant Sahani, and Shwetak Patel. Demographic bias of expert-level vision-language foundation models in medical imaging. *Science Advances*, 11(13):eadq0305, March 2025. doi: 10.1126/sciadv.adq0305. URL https://www.science.org/doi/10.1126/sciadv.adq0305.

Yousef Yeganeh, Azade Farshad, Ioannis Charisiadis, Marta Hasny, Martin Hartenberger, Björn Ommer, Nassir Navab, and Ehsan Adeli. Latent Drifting in Diffusion Models for Counterfactual Medical Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7685–7695, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yeganeh_Latent_Drifting_in_Diffusion_Models_for_Counterfactual_Medical_Image_Synthesis_CVPR_2025_paper.html.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest X-ray radiology report generation. *Patterns*, 4(9), September 2023. ISSN 2666-3899. doi: 10.1016/j.patter.2023.100802. URL https://www.cell.com/patterns/abstract/S2666-3899(23)00157-5.

Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. MEDFAIR: Benchmarking Fairness for Medical Imaging. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=6ve2CkeQe5S. Spotlight.