Measuring What VLMs Don't Say: Validation Metrics Hide Clinical Terminology Erasure in Radiology Report Generation

Aditya Parikh, Aasa Feragen, Sneha Das, and Stella Frank

DTU Compute, Technical University of Denmark

Abstract. Reliable deployment of Vision-Language Models (VLMs) in radiology requires validation metrics that go beyond surface-level text similarity to ensure clinical fidelity and demographic fairness. This paper investigates a critical blind spot in current model evaluation: the use of decoding strategies that lead to high aggregate token-overlap scores despite succumbing to **template collapse**, in which models generate only repetitive, safe generic text and omit clinical terminology. Unaddressed, this blind spot can lead to **metric gaming**, where models that perform well on benchmarks prove clinically uninformative. Instead, we advocate for lexical diversity measures to check model generations for clinical specificity. We introduce Clinical Association Displacement (CAD), a vocabulary-level framework that quantifies shifts in demographic-based word associations in generated reports. Weighted Association Erasure (WAE) aggregates these shifts to measure the clinical signal loss across demographic groups. We show that deterministic decoding produces high levels of semantic erasure, while stochastic sampling generates diverse outputs but risks introducing new bias, motivating a fundamental rethink of how "optimal" reporting is defined. Code available at [19,20].

Keywords: Radiology Report Generation · Validation · Metrics · Fairness · Bias · Semantic Erasure · VLMs · ReXGradient

1 Introduction and Related Work

Radiology report generation (RRG) aims to automate the conversion of medical images into clinically actionable text, reduce documentation burden, and support diagnostic decision-making [5,23]. While recent VLMs have made RRG more viable in clinical workflows [16], in this paper we demonstrate that there is a major gap between scoring well on standard text similarity metrics and providing better clinical reporting or more equitable outcomes across patient populations.

Metrics That Enable Gaming. Contemporary RRG evaluation relies on word-overlap metrics like BLEU [18], ROUGE [13], BERTScore [28], measuring aggregate agreement between generated and reference (ground-truth) reports. Models evaluated on such metrics are vulnerable to **template collapse**: generating high-frequency "normal finding" templates that maximise token overlap despite omitting patient-specific details [26]. As normal radiographs dominate

datasets, such generic outputs can achieve deceptively high scores with low diagnostic value [7]. Standard metrics **cannot detect terminology redistribution**, in which group-specific clinical terms are suppressed, over-generated, or reassigned across demographic groups. The risk increases with VLMs' documented tendency to under-utilize visual evidence in favor of language priors [10], which can lead to more generic outputs reproducing demographic biases. Current evaluation protocols focus on aggregate accuracy, leaving such biases undetected.

Decoding as a Confounding Variable. The inference strategy (deterministic/stochastic decoding) directly shapes the diversity and specificity of generated reports [22]. Deterministic decoding optimises for mode selection, converging to a small set of high-probability templates; stochastic decoding introduces controlled randomness to explore lower-probability tokens. Prior work [12] has motivated stochastic decoding based on marginal gains in BLEU and ROUGE scores. However, selecting decoding strategies solely on average-metric improvements obscures critical fairness trade-offs: while stochasticity may reduce template collapse, it can simultaneously introduce or amplify demographic associations that were absent in reference reports [22]. Existing evaluation frameworks provide no mechanism to audit these lexical shifts.

The Need for Complementary Metrics. Recent work has called for evaluation methods that go beyond surface similarity to measure clinical utility, demographic equity, and semantic fidelity [6]. Clinician-aligned scoring (e.g., RaTEScore [30]) offers one avenue by incorporating expert judgment, but it does not directly quantify how models redistribute clinical terminology across demographic groups. Fairness metrics in NLP often rely on group-wise F1 gaps or counterfactual perturbation tests. What is missing is a **vocabulary-level diagnostic** that explicitly tracks whether the demographic association structure of clinical language is preserved, erased, or distorted during generation.

Our Contributions. (i) We propose **Clinical Association Displacement**, a vocabulary-level framework quantifying shifts in group-associated word usage between human-authored and model-generated reports, identifying erased terminology as well as emergent and preserved bias; (ii) We introduce **Weighted Association Erasure**, a summary statistic that measures the global magnitude of association displacement, diagnosing whether a model systematically avoids clinically informative language; (iii) Using **our radiology VLM ReX-QWEN** as a testbed, we conduct an inference sensitivity analysis, showing that deterministic decoding inflates scores via *template collapse* (high erasure) while stochastic decoding recovers vocabulary but may increase bias; (iv) our baseline comparisons reveal fairness failures invisible to conventional metrics.

Our work establishes that the choice of decoding strategy and metric jointly confounds clinical evaluation. By auditing both what *models say* and systematically *don't say*, we enable transparent evaluation of whether RRG progress reflects genuine clinical utility or evaluation artifacts.

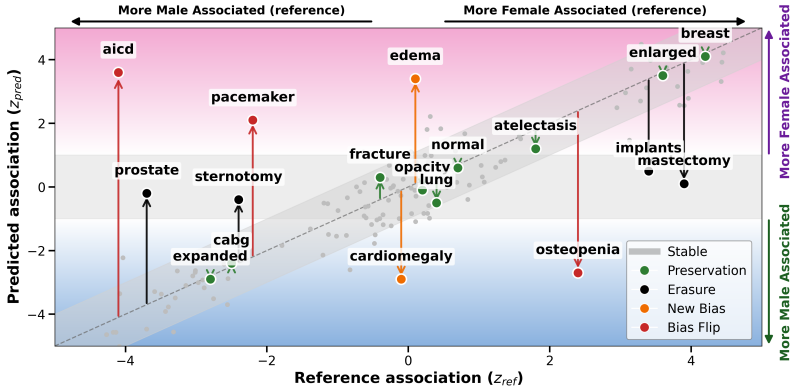

Fig. 1: **Clinical interpretation of CAD.** Each point is a vocabulary term, mapped from its demographic association in the reference corpus (z_{ref}, x-axis) to its association in predictions (z_{pred}, y-axis). Points near the diagonal (gray band) indicate stable associations, while off-diagonal points highlight terms whose demographic association has shifted, reflecting erasure, new bias, or bias flips.

2 Methods

We introduce a corpus-level diagnostic of global lexical shift, quantifying how vocabulary is preserved, erased, or newly introduced during generation. Such shifts may involve clinically meaningful terms that are differentially associated with demographic groups and therefore relevant for diagnostic interpretation. We therefore also introduce diagnostics for demographic-specific shifts, exposing shifts in lexical usage that are not captured by aggregate metrics. We formalize this through: Clinical Association Displacement (CAD) (Section 2.1) and Weighted Association Erasure (WAE) (Section 2.2).

2.1 Clinical Association Displacement (CAD)

CAD is a vocabulary-level diagnostic framework that quantifies how the demographic association, both in polarity and magnitude, of individual clinical terms shifts between a human-reference corpus and generated outputs (see Fig. 1).

Problem Formulation and Log Ratio: Let $\mathcal{D} \in \{\text{ref}, \text{pred}\}$ denote the reference and prediction corpora, where each report is associated with a binary group, here sex: $g \in \{F, M\}$. For each corpus $\mathcal{D}$, we compute, for each word w in the vocabulary $\mathcal{V}$, token counts $c_g^{(\mathcal{D})}(w)$ counting the occurrences of w in reports with sex g. The total token count per group is $N_g^{(\mathcal{D})} = \sum_{w \in \mathcal{V}} c_g^{(\mathcal{D})}(w)$.

To measure how strongly a word w is associated with female vs. male reports in a corpus $\mathcal{D}$, we use a Dirichlet-smoothed log ratio of unigram probabilities [15]. Let $\alpha > 0$ (default: $\alpha = 0.1$) be a smoothing parameter (where $\tilde{}$ denotes α-smoothed estimates) and $V = |\mathcal{V}|$ the vocabulary size. We define:

$$\forall g \in \{F, M\}: \quad p_g^{(\mathcal{D})}(w) = \frac{c_g^{(\mathcal{D})}(w)+\alpha}{N_g^{(\mathcal{D})}+\alpha V} = \frac{\tilde{c}_g^{(\mathcal{D})}(w)}{\sum_{w' \in \mathcal{V}} \tilde{c}_g^{(\mathcal{D})}(w')}, \quad s^{(\mathcal{D})}(w) = \log \frac{p_F^{(\mathcal{D})}(w)}{p_M^{(\mathcal{D})}(w)}.$$

$s^{(\mathcal{D})}(w) > 0$ indicates a *female* association for w, while $s^{(\mathcal{D})}(w) < 0$ indicates that w is more associated to *male*. We apply delta-method approximation under a multinomial model to estimate the variance of $s^{(\mathcal{D})}(w)$:

$$\tilde{\mathrm{Var}}[s^{(\mathcal{D})}(w)] \approx \frac{1}{\tilde{c}_F^{(\mathcal{D})}(w)} + \frac{1}{\tilde{c}_M^{(\mathcal{D})}(w)}, \quad z^{(\mathcal{D})}(w) = \frac{s^{(\mathcal{D})}(w)}{\sqrt{\tilde{\mathrm{Var}}[s^{(\mathcal{D})}(w)]}},$$

yielding a standardized association $z^{(\mathcal{D})}(w)$ adjusting for frequency variance. We denote $s^{(\mathrm{ref})}(w)$, $s^{(\mathrm{pred})}(w)$ and $z^{(\mathrm{ref})}(w)$, $z^{(\mathrm{pred})}(w)$ for reference and predictions.

Quantifying Displacement: We define **Association Displacement** $z_{\mathrm{disp}}(w)$ as the standardized difference between prediction and reference associations, and a critical threshold z_{crit} for significance testing:

$$z_{\mathrm{disp}}(w) = \frac{s^{(\mathrm{pred})}(w) - s^{(\mathrm{ref})}(w)}{\sqrt{\tilde{\mathrm{Var}}[s^{(\mathrm{pred})}(w)] + \tilde{\mathrm{Var}}[s^{(\mathrm{ref})}(w)]}}, \quad z_{\mathrm{crit}} = \Phi^{-1}\left(1 - \frac{p^\star}{2}\right).$$

Assuming $\mathcal{D} \in \{\mathrm{ref}, \mathrm{pred}\}$ are independently identically distributed (IID), the null hypothesis is $z_{\mathrm{disp}}(w) \sim \mathcal{N}(0, 1)$. We convert $z_{\mathrm{disp}}(w)$ to two-sided p-values, apply Benjamini–Hochberg FDR correction [2], and flag words with $|z_{\mathrm{disp}}(w)| > z_{\mathrm{crit}}$ at significance level $p^\star$ (default: $p^\star = 0.05$), where Φ^{-1} is the standard normal quantile function.

Clinical Categories of Shift: For each vocabulary term w we compute $z^{(\mathrm{ref})}(w)$, $z^{(\mathrm{pred})}(w)$ and $z_{\mathrm{disp}}(w)$. Words with $|z_{\mathrm{disp}}| < |z_{\mathrm{crit}}|$ are deemed **stable** (no-displacement) and excluded from further audit. Displaced terms fall into three categories: **(i) Erasure:** A word is strongly associated in the reference ($|z^{(\mathrm{ref})}| > z_{\mathrm{strong}}$) but becomes weak, neutral or is not generated ($|z^{(\mathrm{pred})}| < z_{\mathrm{neutral}}$), indicating loss of clinical signal. **(ii) Emergent Bias:** The model introduces group-associations absent in the reference (**New Bias**: $|z^{(\mathrm{ref})}| < z_{\mathrm{neutral}},\ |z^{(\mathrm{pred})}| \geq z_{\mathrm{neutral}}$) or reverses polarity (**Bias Flip**: $\mathrm{sign}(z^{(\mathrm{ref})}) \neq \mathrm{sign}(z^{(\mathrm{pred})})$, both non-neutral), indicating hallucinated demographic correlations. **(iii) Preservation:** Polarity is conserved ($\mathrm{sign}(z^{(\mathrm{ref})}) = \mathrm{sign}(z^{(\mathrm{pred})})$) without loss of demographic association.

We set $z_{\mathrm{neutral}} = 1$ & $z_{\mathrm{strong}} = 2$ (corresponding to two-tailed normal cutoffs of $p \approx 0.32$ & $p \approx 0.05$), so only statistically reliable deviations from neutrality are treated as clinically meaningful associations.

2.2 Weighted Association Erasure (WAE)

While CAD tracks displacement for individual words, WAE aggregates these term-level shifts to provide a single, global measure of clinical signal loss across the entire vocabulary, weighted by clinical relevance.

$$\mathrm{WAE} = \frac{\sum_{w \in \mathcal{V}} \omega(w) \cdot z_{\mathrm{disp}}(w)^2}{\sum_{w \in \mathcal{V}} \omega(w)}$$

where the numerator represents the *aggregate displacement energy* scaled by a term importance factor $\omega(w)$ (approximated by term frequency). The denominator is normalization factor to adjust for the total number of predicted tokens, enabling WAE to be compared across output corpora of different lengths.

We define the weight $\omega(w) = c^{(\mathrm{pred})}(w)$ (prediction-based) or $\omega(w) = c^{(\mathrm{ref})}(w)$ (reference-based). Prediction-based weighting quantifies shifts within text the model *actually produces*, avoiding penalties for omitted vocabulary. Reference-based weighting reveals which clinically important terms (frequent in human reports) are systematically avoided. Aggregating over terms associated with group g in the reference yields a group-specific score WAE_g, and disparity is defined as ΔWAE (e.g., $\Delta\mathrm{WAE} = \Delta\mathrm{WAE}_F - \Delta\mathrm{WAE}_M$).

Uncertainty is estimated via nonparametric bootstrap; statistical calibration is confirmed via label-permutation tests, where observed WAE consistently exceeds the null distribution under random demographic assignment.

3 Experiments

Dataset: We use **ReXGradient-160K**, the largest publicly available multi-site chest X-ray dataset [29], which includes clinician-authored findings and impressions. We filter frontal views using a foundational model [14] and create a stratified test-subset of $N = 2,000$ with balanced sex representation.

Models: ReX-QWEN: We fine-tune Qwen3-VL-8B-Instruct [25] on ReXGradient using LoRA [9](rank $r = 16$, $\alpha = 32$) targeting attention and MLP layers. Input images are resized to 512×512 RGB. Our primary model, **ReX-BB** (Balanced Baseline) is trained on 60K sex-balanced examples (50% male, 50% female) as our primary model. Two sex-skewed variants: ReX-MB (67% male) and ReX-FB (67% female) validate CAD/WAE sensitivity to demographic shifts. We compare ReX-QWEN against three models under identical evaluation protocols: **CXR-LLaVA** [11], **LLaVA-Rad** [27] (medical VLMs for radiology reporting); and **OpenAI GPT-4.1** [17], prompted zero-shot for report generation.

Inference Strategy Study: We hypothesise that semantic erasure stems from inference strategies that systematically suppress clinical terminology to minimise generation risk. We partition inference strategies into **deterministic** (greedy, beam search), which select the highest-probability sequences and may lead to template collapse, and **stochastic** (temperature T, top-k, nucleus sampling), which sample from truncated distributions to increase lexical diversity.

Specifically, we evaluate: (i) Greedy (argmax), (ii) Beam Search ($k = 4$), (iii) Conservative (Low) Sampling $T = 0.3$, top-$p = 0.95$), (iv) Rich Sampling ($T = 0.7$, top-$p = 0.95$), (v) Top-k ($k = 50$, $T = 1.0$), and (vi) Nucleus Sampling

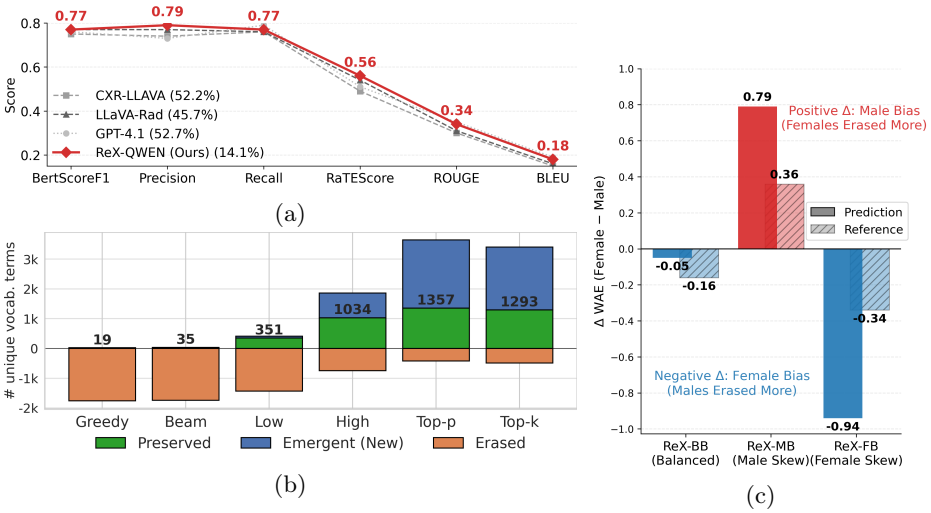

Fig. 2: **(a)** Clinical performance comparison (w/ Template Diversity %). **(b)** Global lexical shift across inference strategies. **(c)** ΔWAE captures semantic shifts caused by sex-skewed training.

(top-$p = 0.9$, $T = 1.0$). All use max tokens=512 with caching enabled.

Evaluation Metrics: We evaluate models across three dimensions: **Clinical Accuracy:** We report BERTScore (F1, Precision, Recall) using CheXbert as the backbone model [24], ROUGE-1/L, and RateScore to measure semantic and surface-level similarity against reference reports. **Textual Diversity:** Following [8], we compute: (i) Template Diversity (fraction of unique reports), (ii) Type-Token Ratio (TTR), (iii) 1-Self-BLEU, and (iv) Semantic Diversity. **Our Metrics:** We report: (i) our proposed CAD-flagged clinical categories (Erasure, New Bias, Bias Flip) quantifying vocabulary-level association shifts, (ii) our proposed WAE summarizing global displacement magnitude, (iii) the existing WEAT [3,1] and its effect size (ES) as a comparative word-association baseline, and (iv) Disparity Impact Ratio (ΔDIR) [21], the change in male-to-female token ratio between reference and prediction. Stop words are removed prior to all association-based metric computation to ensure focus on clinical terms.

4 Results and Discussion

We establish that while ReX-QWEN achieves competitive baseline performance (see Fig. 2a), severe template collapse remains evident (Template Diversity 14.1−52.7%). This collapse motivates our subsequent lexical analysis, specifically investigating how decoding choices drive a fundamental trade-off between semantic erasure and emergent bias. To evaluate these dynamics, we compare ReX-BB (sex-balanced) across six distinct inference strategies (Table 1).

Table 1: **Inference strategy sensitivity.** (**Bold**: best; Underline: worst).

Metric Category	Ref.	Deterministic		Stochastic (Increasing Entropy)			
		Greedy	**Beam**	$T = 0.3$	$T = 0.7$	**Top-k**	**Top-p**
Surface Level Similarity							
BERTScore F1 ↑	–	**0.78**	**0.78**	0.76	0.73	0.71	0.69
BERTScore P/R ↑	–	0.80/0.75	**0.81/0.76**	0.78/0.75	0.73/0.74	0.68/0.74	0.70/0.73
ROUGE (1/L) ↑	–	0.34/0.29	**0.35/0.30**	0.32/0.26	0.26/0.19	0.20/0.14	0.21/0.15
RateScore ↑	–	**0.55**	**0.55**	0.54	0.48	0.44	0.44
Textual Diversity & Collapse							
Template Div. (%) ↑	82.45	0.3	0.5	36.3	97.05	**100.0**	**100.0**
TTR (%) ↑	2.915	0.006	0.131	0.654	2.128	2.540	**2.781**
1 - Self-BLEU (%) ↑	35.70	1.05	0.20	10.78	33.78	46.73	**46.84**
Semantic Div. (%) ↑	85.23	12.16	5.41	55.51	83.75	85.80	**86.19**
Clinical Categories (Ours)							
Erasure (#) ↓	–	32	35	17	4	5	**3**
New Bias (#) ↓	–	**0**	3	41	45	25	65
Bias Flip (#) ↓	–	**0**	1	16	33	44	69
Fairness							
WAE (pred) (Ours) ↓	–	**2.49**	2.50	4.95	33.81	12.90	13.27
WAE (ref) (Ours) ↓	–	**0.19**	0.60	3.43	9.06	4.60	14.70
WEAT (S/ES) ↓	–	-1.46/-0.64	1.10/0.03	-1.35/-0.01	4.30/-0.07	-2.13/-0.02	0.27/0.00
ΔDIR ↓	–	-0.03	**0.01**	-0.08	-0.01	-0.20	-0.32

Global lexical collapse. Deterministic strategies exhibit severe lexical attrition, generating vocabularies of just 22–38 unique terms, 5% of the original reference vocabulary (see Fig. 2b), collapsing instead to "normal-finding" templates. Stochastic approaches preserve vocabulary size but introduce emergent terms absent in references, raising validity concerns: are these recovered clinical signals or hallucinated noise? To determine if these shifts disproportionately affect specific patient groups, we turn to CAD to audit for semantic erasure and emergent bias in sex-associated terms.

Deterministic decoding inflates similarity via collapse. Greedy and beam search produce 0.3–0.5% unique reports, presenting a case of *template collapse*. Despite BERTScore F1=0.78, CAD reveals severe sex-association erasure: 32–35 erased terms, corresponding to nearly all the highly-biased terms in the references. For example, *AICD*, *prostate* (male-associated), and *vasculature*, *pregnancy*, *mastectomy* (strongly female-associated) shift their predicted association to neutral. This metric-gaming artifact confirms that high aggregate scores can mask systematic lexical erasure of terminology present in reference reports.

Stochastic decoding restores vocabulary but introduces bias. Temperature scaling increases diversity (36.3–100%) and reduces erasure but introduces 25–65 new bias terms and elevates WAE_{pred} (2.50 → 33.81) due to the emergence of newly biased terms. ΔDIR widens, signaling verbosity imbalances across sex

Table 2: **External validation confirms the persistence of the erasure-versus-bias trade-off using LLaVA-Rad with different decoding strategies** Text similarity metrics (BERTScore, ROUGE etc.) are comparable ($\Delta \leq 2\%$, unshown), while Diversity metrics show clear differences in template usage (Div. Temp.) and erasure rates. (NB: New Bias; BF: Bias Flip)

Metrics	ReXGradient		CheXpert	
	Greedy	Rich	Greedy	Rich
Div. (Temp./Sem.%) ↑	45.70/70.32	91.05/79.49	94.05/80.60	100/83.13
Erasure/NB/BF ↓	19/6/5	11/5/11	6/1/0	1/1/4
WAE (pred/ref) ↓	1.17/0.84	1.03/0.93	0.76/0.60	0.84/0.69
WEAT (S/ES) ↓	1.13/0.01	-0.36/-0.00	-0.78/0.11	-0.14/0.17

groups. For instance, the pathology *pneumothorax* is neutral in human-reference ($z_{ref} = 0.15$); however low temperature generation actively hallucinated a strong association with male patients ($z_{pred} = -5.69$); more examples of induced biases: *calcification* (+1.12 → -0.24), *pectoralis* (-0.17 → 2.88), *asthma* (-2.96 → 2.54), *contours* (+7.47 → -3.79), *hyperinflation* (-0.22 → -3.77), *fractures* (-8.04 → 1.93). WEAT evaluates static latent embeddings and therefore remains entirely blind to decoding-induced biases.

Implications. CAD identifies two opposing failure modes: **(i) semantic erasure** under deterministic decoding: systematic vocabulary suppression invisible to BERTScore/ROUGE; **(ii) emergent bias** under high-temperature sampling: new demographic associations absent in references. No single strategy simultaneously optimises accuracy, diversity, and fairness, motivating fairness-aware decoding selection tuned jointly on clinical accuracy *and* CAD/WAE diagnostics rather than defaulting to greedy inference. **We recommend** augmenting standard evaluation with CAD categories, WAE, and lexical diversity metrics to prevent interpreting template-collapse artifacts as genuine clinical improvement.

Generalisation and Robustness: LLaVA-Rad on ReXGradient and CheXpert [4] confirms the pattern (Table 2): deterministic decoding reduces template diversity (45.7% vs. 91.1%) with higher erasure (19 terms), while stochastic sampling introduces bias flips (5→12). WAE drops modestly (1.17→1.03 on ReXGradient), confirming the trade-off persists across architectures and datasets. Sex-skewed training (ReX-MB, ReX-FB) leaves conventional metrics stable (F1 (BertScore)$\in [0.76, 0.78]$, RateScore$\in [0.54, 0.57]$, WEAT-ES≈ 0). However, $\Delta\text{WAE}_{\text{pred}}$ goes from near-zero (ReX-BB: -0.05) to +0.79 for ReX-MB and −0.94 for ReX-FB, confirming WAE captures semantic shifts invisible to existing metrics (refer Fig. 2c). **For Ablation**, we sweep $p^{\star}_{\text{disp}}$ across its full range strictly preserving the relative ranking of all decoding strategies. This ablation confirms that the erasure-versus-bias trade-off is a fundamental generation mechanic, rather than an artifact of metric thresholding.

5 Conclusion and Future Work

We demonstrate that decoding strategy and metric choice jointly confound evaluation in radiology report generation. By introducing CAD and WAE, we provide a framework to audit semantic erasure and emergent bias that remain completely invisible to conventional metrics. Our findings are diagnostic at the corpus and lexical level. Together, these tools enable transparent evaluation, allowing the field to distinguish genuine clinical utility from the illusion of performance caused by *template collapse*. **Future Work:** Our analysis focuses on binary sex associations; future work should extend CAD to age, ethnicity, and intersectional groups. Clinician validation is needed to confirm whether CAD-flagged erasures correspond to clinically meaningful omissions.

Disclosure of Interests

The authors have no competing interests to declare.

Acknowledgment

This work was funded by the Novo Nordisk Foundation (project number 0087102).

References

1. Badilla, P., Bravo-Marquez, F., Zambrano, M.J., Pérez, J.: Wefe: A python library for measuring and mitigating bias in word embeddings. Journal of Machine Learning Research **26**(156), 1–6 (2025), `http://jmlr.org/papers/v26/22-1133.html`
2. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) **57**(1), 289–300 (1995)
3. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017)
4. Chambon, P., Delbrouck, J.B., Sounack, T., Huang, S.C., Chen, Z., Varma, M., Truong, S.Q., Chuong, C.T., Langlotz, C.P.: Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. arXiv preprint arXiv:2405.19538 (2024)
5. Demner-Fushman, D., Chapman, W.W., McDonald, C.J.: What can natural language processing do for clinical decision support? Journal of biomedical informatics **42**(5), 760–772 (2009)
6. Du, X., Zhou, Z., Wang, Y., Chuang, Y.W., Li, Y., Yang, R., Zhang, W., Wang, X., Chen, X., Guan, H., et al.: Testing and evaluation of generative large language models in electronic health record applications: a systematic review. Journal of the American Medical Informatics Association p. ocaf233 (2026)
7. Gao, L., Zhang, L., Liu, C., Wu, S.: Handling imbalanced medical image data: A deep-learning-based one-class classification approach. Artificial intelligence in medicine **108**, 101935 (2020)
8. Guo, Y., Shang, G., Vazirgiannis, M., Clavel, C.: The curious decline of linguistic diversity: Training language models on synthetic text. In: Findings of the Association for Computational Linguistics: NAACL 2024. pp. 3589–3604 (2024)

9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
10. Lee, K.i., Kim, M., Yoon, S., Kim, M., Lee, D., Koh, H., Jung, K.: Vlind-bench: Measuring language priors in large vision-language models. In: Findings of the Association for Computational Linguistics: NAACL 2025. pp. 4129–4144 (2025)
11. Lee, S., Youn, J., Kim, H., Kim, M., Yoon, S.H.: Cxr-llava: a multimodal large language model for interpreting chest x-ray images. European Radiology **35**(7), 4374–4386 (2025)
12. Li, Q.: Harnessing the power of pre-trained vision-language models for efficient medical report generation. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. p. 1308–1317. CIKM '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3583780.3614961, https://doi.org/10.1145/3583780.3614961
13. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
14. Ma, D., Pang, J., Gotway, M.B., Liang, J.: A fully open ai foundation model applied to chest radiography. Nature **643**(8071), 488–498 (2025)
15. Monroe, B.L., Colaresi, M.P., Quinn, K.M.: Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. Political Analysis **16**(4), 372–403 (2008)
16. Nam, Y., Kim, D.Y., Kyung, S., Seo, J., Song, J.M., Kwon, J., Kim, J., Jo, W., Park, H., Sung, J., et al.: Multimodal large language models in medical imaging: current state and future directions. Korean Journal of Radiology **26**(10), 900 (2025)
17. OpenAI: Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/ (2025), accessed: 2026-02-24
18. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
19. Parikh, A.: clinical-cad: Clinical association displacement (2026), https://pypi.org/project/clinical-cad/
20. Parikh, A.: ReX-QWEN: Radiology vlm for clinical association displacement framework (2026), https://github.com/ADE-17/ReX-QWEN---Radiology-VLM-for-Clinical-Association-Displacement
21. Parikh, A., Das, S., Feragen, A.: Investigating label bias and representational sources of age-related disparities in medical segmentation. arXiv preprint arXiv:2511.00477 (2025)
22. Presacan, O., Nik, A., Thambawita, V., Ionescu, B., Riegler, M.: A comparative study of decoding strategies in medical text generation. In: International Conference on Multimedia Modeling. pp. 1–15. Springer (2026)
23. Rocha, D.M., Brasil, L.M., Lamas, J.M., Luz, G.V., Bacelar, S.S.: Evidence of the benefits, advantages and potentialities of the structured radiological report: An integrative review. Artificial intelligence in medicine **102**, 101770 (2020)
24. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A.Y., Lungren, M.P.: Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert (2020)
25. Team, Q.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388
26. Yun, L., An, C., Wang, Z., Peng, L., Shang, J.: The price of format: Diversity collapse in llms. arXiv preprint arXiv:2505.18949 (2025)

27. Zambrano Chaves, J.M., Huang, S.C., Xu, Y., Xu, H., Usuyama, N., Zhang, S., Wang, F., Xie, Y., Khademi, M., Yang, Z., et al.: A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. Nature Communications **16**(1), 3108 (2025)
28. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)
29. Zhang, X., Acosta, J.N., Miller, J., Huang, O., Rajpurkar, P.: Rexgradient-160k: A large-scale publicly available dataset of chest radiographs with free-text reports. arXiv preprint arXiv:2505.00228 (2025)
30. Zhao, W., Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Ratescore: A metric for radiology report generation. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 15004–15019 (2024)