Universal Boosts, Specific Suppressors: Sparse Autoencoder Steering of Medical Vision-Language Models

Anonymous ACL submission

Abstract

Medical vision-language models (VLMs) often hallucinate findings when generating chest X-ray reports: they fabricate findings that are not present in the image, miss important ones, or locate them incorrectly. We mitigate this *without* weight updates by decoding-time residual steering on a per-token sparse autoencoder (SAE) basis: Top-K SAEs on late layers, causal steering against clinical errors, then combined *suppress/boost* intervention at inference time. On the MIMIC-CXR test split, our inference-only method improves the quality of generated reports for three radiology VLMs (RadVLM, LLaVA-Rad, and CheXOne), with relative improvements of +5.4%, +7.2%, and +17.0% in the clinical composite metric, and statistically significant GREEN gains on all backbones. A cross-model feature alignment shows that the quality-promoting (*boost*) directions overlap strongly across architectures, whereas hallucination-linked (*suppress*) directions are model-specific. Therefore, transferable steering must treat suppression per-backbone, rather than sharing a universal suppress list. The same recipe transfers zero-shot to IU-Xray (Green $+7.7\%$ rel.) without retraining, confirming that the identified features are properties of the model, not of the training corpus. The causal feature sets and an interactive feature dashboard are enclosed as supplementary material.

1 Introduction

Medical vision-language models (VLMs) can hallucinate findings, omit pathologies, or mislocalize observations during the generation of radiology reports from medical images (Calamida et al., 2023; Ostmeier et al., 2024). The most common approach to mitigate this problem is post training-time optimization using Reinforcement Learning (RL) by employing Direct Preference Optimization (DPO) or Group Relative Policy Optimization (GRPO) fine-tuning on medical VLMS, such as CheXalign (Hein et al., 2025), RadVLM-GRPO (Gundersen et al., 2026), and CheXOne (Zhang et al., 2026a), which achieved strong gains. However, these remedies require gradient updates, optimized reward signals, and per-model training runs, which result in substantial computational overhead (e.g., ∼350 H200 GPU-hours for RadVLM-GRPO; Gundersen et al., 2026).

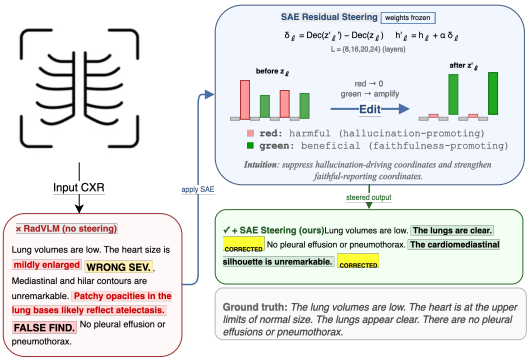


Figure 1: **SAE hallucination mitigation on a real CXR (Schematic illustration).** The unsteered RadVLM hallucinates WRONG SEVERITY and FALSE FINDING. Each SAE feature is classified as harmful or beneficial by a *prior causal screen*: zeroing the feature on a validation set and measuring whether clinical errors (GREEN) increase or decrease (§3). In inference, harmful features are zeroed, beneficial ones amplified, and the resulting residual $\boldsymbol{\delta}$ is added to the frozen hidden state. The steered report matches ground truth.

Here, we take an orthogonal approach, inspired by Sparse Autoencoders (SAEs) (Bricken et al., 2023), by freezing the model and *editing the model's internal activations at inference time*. SAEs decompose transformer hidden states into high-dimensional sparse representations (Bricken et al., 2023; Huben et al., 2024; Gao et al., 2025) and then produce interpretable features that could be linked to specific model behaviors. Recent work shows that SAE features are also useful for steering when the proper features are selected (Arad et al., 2025; O'Brien et al., 2025). We apply these find-

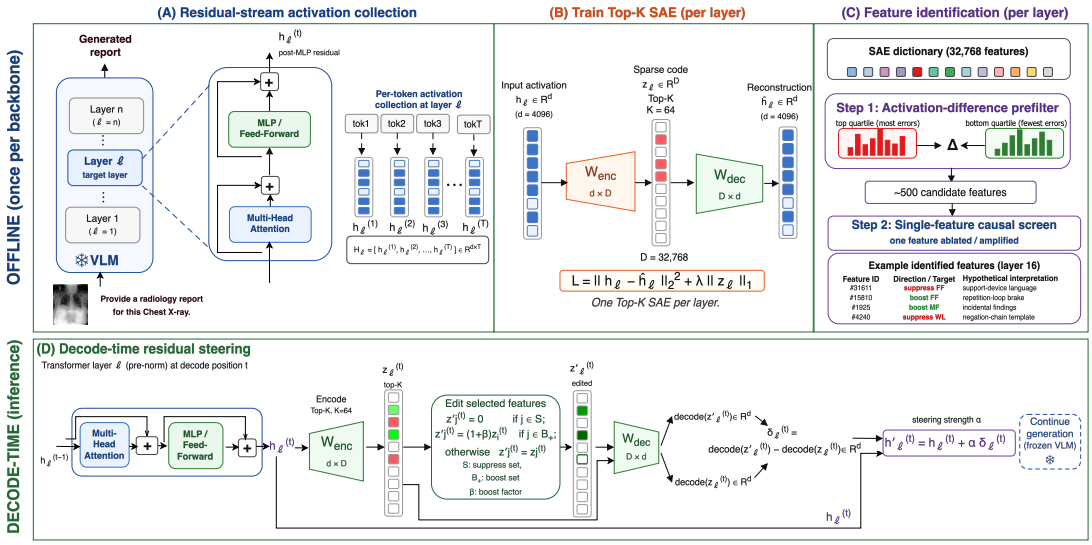


Figure 2: End-to-end SAE residual-steering pipeline: (weights frozen).**(A)** Collect per-token residual activations $\mathbf{h}_\ell^{(t)}$ while the VLM generates reports. **(B)** Train a layer-wise Top-K SAE to obtain sparse codes $\mathbf{z}_\ell$.**(C)** Identify which coordinates to edit. Validation studies are ranked by GREEN error count and split into quartiles. *Step 1:* keep $\sim$500 features with the largest $|\bar{z}_j^{\text{high}} - \bar{z}_j^{\text{low}}|$, i.e. the mean-activation gap between the top quartile (most errors) and bottom quartile (fewest errors) *Step 2:* for each candidate, ablate it on held-out decodes and measure the change in per-type GREEN errors; ablation that *lowers* errors $\Rightarrow$ suppress, ablation that *raises* errors $\Rightarrow$ boost. **(D)** At inference, apply those edits to $\mathbf{z}_\ell^{(t)}$ (zero suppress, scale boost), decode, and add only the residual difference $\boldsymbol{\delta}_\ell$ to $\mathbf{h}_\ell^{(t)}$ with strength α.

ings to the safety-critical domain of medical radiograph report generation, where the interpretability of the machine learning model is important. We organize this work around four research questions that examine the presence (Q1), causal relevance (Q2), generalizability (Q3), and structure of clinically meaningful SAE features (Q4) in medical radiograph VLMs as follows:

Q1: Do per-token Top-K SAEs trained on medical VLMs residual streams uncover a small, sparse set of features whose activation has a clinically meaningful effect on *each* clinical error type (false findings, missing findings, wrong location, wrong severity)?

Q2: Can such features be used to mitigate hallucinations at *inference time only*, by editing the residual stream in the SAE basis, without any weight update, or change to the prompt or the decoding algorithm?

Q3: Does the same identification and steering recipe transfer across three medical VLM backbones that differ in size, vision encoder, and training regime, and does it remain effective *on top of* a reinforcement-learning fine-tune? Does it also transfer *across datasets*, from MIMIC-CXR (where features are identified) to the independently collected IU-Xray, with no re-training?

Q4: Looking across these three architectures, do they share the same hallucination-mitigating mechanisms, or do quality-promoting and hallucination-linked directions decompose differently?

2 Related work

Sparse autoencoders for interpretability. Dictionary learning over transformer activations recovers sparse, interpretable features (Bricken et al., 2023; Huben et al., 2024; Gao et al., 2025; Templeton et al., 2024). Marks et al. (2025) by utilizing single-feature interventions to trace causal circuits at the level of individual SAE units and exploring which circuit a feature participates in.

We adopt the single-feature intervention approach to answer a different question: *which clinical error type does an SAE feature causally control?*

Inference-time activation editing. Activation addition (Postmus and Abreu, 2024), representation engineering (Andy Zou, 2023), and inference-time intervention (ITI) (Li et al., 2023) all edit hidden states without changing weights. SAE-based steering refines this line of work by operating in a sparse, interpretable basis: Arad et al. (2025) shows that the choice of feature is decisive, and O'Brien et al.

(2025) applies SAE steering to control refusal in language models.

We extend the SAE-steering line into a structured clinical setting with multiple error types and multi-metric evaluation, making the cross-architecture census the centerpiece rather than a side observation.

SAE steering of vision-language models. Pach et al. (2026) train SAEs on the *vision encoder* of CLIP and erase visual concepts (e.g., *knives*, *laptops*) in LLaVA's generated descriptions by single-neuron suppression. Their intervention comes before ours in the process; we are the language-stream counterpart in a clinical setting, residual edits on the LLM, structured clinical error types, and feature selection by causal screening against GREEN (Ostmeier et al., 2024) rather than visual monosemanticity.

Hallucination in medical VLMs. Most remedies are applied at training-time such as (a) supervised fine-tuning on curated reports (LLaVA-Rad Zambrano Chaves et al. (2025); RadVLM (Deperrois et al., 2025b); MAIRA-2 (Bannur et al., 2024)), (b) clinical RL with reward engineering (CheXOne (Zhang et al., 2026a); RadVLM-GRPO (Gundersen et al., 2026)), or (c) region-guided grounding (RGRG (Tanida et al., 2023)); these approaches require gradient updates and model-specific rewards, grounding, or fine-tuning signals. Other alternatives such as decoding-time interventions, reshape next-token logits rather than hidden states (VCD (Leng et al., 2024); OPERA (Huang et al., 2024); DoLa (Chuang et al., 2024)), but are validated on general-domain benchmarks and do not separate clinical error types. We instead edit the residual stream on an SAE basis: an inference-only method that exposes interpretable per-GREEN-error-type controls, stacks on medical VLMs, and is evaluated with structured clinical metrics.

To our knowledge, this is the **first application of SAE-based steering to medical report generation**, and we further show that a per-model SAE pipeline transfers cleanly onto a GRPO-tuned medical VLM (CheXOne).

3 Method

Our approach intervenes on a frozen radiology VLM at decoding time and consists of four stages as shown in Figure 2:

A) Residual-stream Activation Collection Let $\mathbf{h}_\ell^{(t)}$ denote the hidden state at transformer layer ℓ and token position t during report generation. We save per-token activation for all layers $\mathcal{L}$, which is used to *train* sparse autoencoders (SAEs), and a set of layers for intervention hooks, across training and validation-split samples. These tensors correspond to the SAE training datasets and the intervention locations.

(B) Sparse Autoencoder Training At each layer ℓ, a Top-K SAE encodes $\mathbf{h}_\ell^t$ into a sparse representation $\mathbf{z}_\ell^{(t)} \in \mathbb{R}^D$ with at most K nonzeros and decodes $\hat{\mathbf{h}}_\ell^{(t)} = \mathrm{decode}_\ell(\mathbf{z}_\ell^{(t)})$ (Gao et al., 2025). We use dictionary size $D{=}32,768$, sparsity $K{=}64$, and model hidden dimension $d{=}4,096$.

(C) Feature Identification We use two approaches to select features to suppress or amplify.
1. Correlation based selection (Corr) – For each layer ℓ and feature j, we compute the Pearson correlation between the feature's activation magnitude on the training set and the corresponding GREEN error count (overall or per error type). Features with the most positive correlation become candidates for *suppression*; the most negatively correlated become *boost* candidates. The procedure yields per-error-type lists out of the box.
2. Causal selection (Causal) – This involves a two-stage procedure. First, an *activation-difference prefilter*: we keep the top $\sim$500 features exhibiting the largest absolute gap in mean activation between top-quartile (high-error) and bottom-quartile (low-error) samples. Second, a *single-feature causal screen*: for each candidate, we run a one-feature intervention (zero or amplify) on a held-out validation set and measure the change in the automatic clinical metrics directly (Arad et al., 2025). For each surviving feature j, we record the *per-error-type causal data* as:

$$\Delta_t^{(j)} = \frac{1}{N}\sum_{i=1}^{N}\left(\#\mathrm{GREEN}_t(a_i^{(j)}) - \#\mathrm{GREEN}_t(b_i)\right) \quad (1)$$

for $t \in \{\mathrm{FF}, \mathrm{MF}, \mathrm{WL}, \mathrm{WS}\}$, where b_i and $a_i^{(j)}$ are the baseline and feature-j-ablated decodes of validation study i, and $\#\mathrm{GREEN}_t(\cdot)$ is the count of GREEN clinically-significant errors: *false findings* (FF), *missing findings* (MF), *wrong location* (WL), and *wrong severity* (WS) (see § 4). $\Delta_t^{(j)}$ represents the mean errors per report. Features are then ranked by their measured causal effect on each error type, not by association alone.

Both methods produce, per layer, two ordered lists including features to **suppress** ($\Delta_t^{(j)} < 0$ asso-

ciated with errors) and features to **boost** ($\Delta_t^{(j)} > 0$ associated with quality), organized across the four GREEN error types.

(D) Residual Steering Given our identified feature lists, we apply inference-time steering to the model's residual stream. We emphasize a *residual* update rather than standard activation patching to prevent SAE reconstruction error from degrading generation quality.

1. Feature Aggregation & Conflict Resolution – At inference time, we truncate our ranked lists to a validation-chosen budget K. For causal selection, we aggregate the top-K features across all GREEN error types into a master suppress list ($\mathcal{S}_\ell$) and a master boost list ($\mathcal{B}_\ell$). Because a feature may inadvertently appear on both lists due to differing error profiles, we resolve conflicts by strictly prioritizing suppression to minimize clinical hallucinations: any overlapping feature is removed from $\mathcal{B}_\ell$.

2. Feature Editing – Let $\mathbf{z}_\ell$ denote the SAE representation of the hidden state $\mathbf{h}_\ell$. We construct the edited features $\mathbf{z}'_\ell$ by forcefully zeroing out suppressed features and scaling boosted features by a factor of $(1+\beta)$:

$$z'_{\ell,j} = \begin{cases} 0 & \text{if } j \in \mathcal{S}_\ell \\ (1+\beta)\, z_{\ell,j} & \text{if } j \in \mathcal{B}_\ell \\ z_{\ell,j} & \text{otherwise.} \end{cases} \tag{2}$$

In our experiments, $\beta = 1.0$, doubling boosted coordinates, while baseline unsteered sets use $\alpha = 0$ and $\beta = 0$).

3. Applying the Residual Update – Standard SAE-mediated activation patching (Marks et al., 2025; Templeton et al., 2024) replaces the original hidden state with the SAE reconstruction:

$$\mathbf{h}'_\ell \;=\; (1-\alpha)\,\mathbf{h}_\ell + \alpha\,\mathrm{decode}_\ell(\mathbf{z}'_\ell) \tag{3}$$

However, this injects baseline reconstruction error ($\mathrm{decode}_\ell(\mathbf{z}_\ell) - \mathbf{h}_\ell$) directly into the network.

To avoid this, we apply a pure residual update. We isolate the exact delta caused by our edits and add it directly back to the original hidden state, scaled by steering strength α:

$$\boldsymbol{\delta}_\ell = \mathrm{decode}_\ell(\mathbf{z}'_\ell) - \mathrm{decode}_\ell(\mathbf{z}_\ell), \quad \mathbf{h}'_\ell = \mathbf{h}_\ell + \alpha\,\boldsymbol{\delta}_\ell \tag{4}$$

This formulation is the SAE-basis analogue of additive inference-time edits (Li et al., 2023): it exactly cancels out the reconstruction residue and preserves the unedited knowledge in $\mathbf{h}_\ell$. We retain Eq. (3) only as an early ablation (it degrades sharply for $\alpha \geq 0.25$ in our setting) and apply Eq. (4) at every layer in $\mathcal{L}$ for every generated token.

4 Experiments

Dataset and Metrics. We use **MIMIC-CXR-JPG** (Johnson et al., 2024), a large public chest-radiograph dataset paired with radiology reports. Following Deperrois et al. (2025b), we restrict to single frontal chest X-rays paired with findings, where references to prior studies are removed. After filtering there are 230,980 image-text pairs in the training/validation set and 3,314 in the test set. The corresponding datapoints can be extracted from the RadVLM instruction dataset (Deperrois et al., 2025a). SAE training and feature identification use the training split; hyperparameter tuning (α, number of features, steering mode) uses the validation split; test-set numbers are reported once, on a frozen configuration.

To test whether features identified on MIMIC-CXR transfer to an unseen corpus, and to provide DUA-compliant qualitative examples (MIMIC-CXR's data-use agreement prohibits reproducing images and reports in publications), we additionally evaluate on **IU-Xray** (Demner-Fushman et al., 2016) ($N = 3,307$ frontal studies after filtering for image availability). No SAE retraining, feature re-screening, or hyperparameter re-tuning is performed: the same weights, feature lists, and (α, K) from MIMIC-CXR are applied zero-shot. Results are reported in Appendix F.

We primarily employ GREEN (Ostmeier et al., 2024, based on *StanfordAIMI/GREEN-radllama2-7b*), which classified errors into six categories: *False Findings* (FF): False report of a finding in the candidate, *Missing Findings* (MF): Missing a finding present in the reference, *Wrong Location* (WL): Misidentification of a finding's anatomic location/position, *Wrong Severity* (WS): Misassessment of the severity of a finding, *False Comparisons* (FC): Mentioning a comparison that isn't in the reference, and *Missing Comparisons* (MC): Omitting a comparison detailing a change from a prior study.

We focus on FF, MF, WL, and WS since references to prior studies were filtered. We report the GREEN score. Further, we report other clinical metrics `RadGraph-F1 simple` (Jain et al., 2021), `CheXbert micro F1, 14 labels` (Smit et al., 2020), and `BERTScore` (Zhang et al., 2020). We formulate the `Composite Score` as $0.4 \times GREEN + 0.3 \times RadGraph + 0.2 \times CheXbert + 0.1 \times BERTScore$. We also report `RadCliQ` (Yu et al., 2022), a learned index that combines RadGraph,

BERTScore, CheXbert embeddings, and BLEU, calibrated to radiologist preferences. Furthermore, BLEU-4 and ROUGE-L are evaluated as general text generation metrics. These scores are calculated using the RadEval framework (Xu et al., 2025).

Base Models. The primary VLM we use is RadVLM (Qwen-3-VL-8B backbone, Radiology instruction-tuned). All weights in the model remain frozen throughout the pipeline. For generalization aspects, we choose LLaVa-Rad (7B, LLaVa/Vincua) and CheXOne (Qwen2.5-VL-3B, GRPO-tuned). For each mode, we report unsteered test inference and SAE residual steering using SAEs and targeted features trained on that model's own activations, with hyperparameters chosen on the validation set. CheXOne is evaluated in instruct mode (standard findings prompt, no long reasoning trace) for comparability with the other report generation models.

4.1 Experimental Pipeline

We run the end-to-end recipe from § 3 as a sequence of separated jobs: activations collection → SAE training → feature selection → hyperparameter search on validation → generate reports on test set. We also perform the same recipe on LLaVA-Rad and CheXOne.

End-to-end Configuration The following setups define the pipeline we use for all main experiments (RadVLM, unless we note otherwise) on top of the methods explained previously. For SAE training and per-token activation collection, we use a fixed subset of $3,000$ studies drawn *only* from the training split. Indices are selected with *stratified sampling* over CheXpert-style pathology groups (primary label per study, with a minimum count per group) and a fixed random seed, so rare findings are represented rather than a random slice dominated by common classes. We limit the generated reports (decoding length) to 512 tokens, and we train one Top-K SAE per SAE training layer (e.g, ten layers $\{0, 4, 8, 12, 16, 20, 24, 28, 32, 35\}$ for RadVLM's 36-block stack). We attach steering hooks to the four mid/late layers $\mathcal{L}=\{8, 16, 20, 24\}$, where SAE reconstruction quality is high, and activation norms remain moderate enough for steering perturbations to be effective (Appendix A, Table 5, as layer ablations).

Correlation-based selection uses the same 3,000 examples by later activation files paired with per-sample GREEN error count. In the causal selection process, the prefilter runs over all D dictionary features at each of the four hook layers, then the single-feature forward screens on only 100 held-out validation examples. We split validation studies by RadCliQ quality into four quartiles (worst 25%, ..., best 25%), then draw roughly equal counts from each quartile, so that the panel mixes low- and high-quality reports rather than being dominated by one regime. The target features are recorded as per-layer JSON files (top-K indices for suppress and boost, optionally by GREEN type).

Hyperparameter search. For each model backbone after activation collection and SAEs training, we perform a grid search over the subset of validation split (200 stratified samples), over steering strength $\alpha \in \{0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.50\}$, and per-layer feature count $K \in \{20, 50, 100\}$, steering approach (residual vs. blend; Combined vs. suppress-only vs. boost-only), the layer subset/multi-layer ablation $\mathcal{L}$ ($\{8, 16, 20, 24\}$), and the boost factor $\beta \in \{0.5, 1.0\}$. Each condition runs a full report generation on all validation cases with SAE hooks at the four hook at the four steered layers and is scored with GREEN, RadGraph, CheXbert, and BERTScore.

5 Main results

We organize our empirical findings to discuss the design choices incrementally. We first identify the SAE steering operating point per-model by ablating one design factor at a time on the validation split (§ 5.1), then evaluate the chosen configuration on the held-out test set with paired bootstrap significance (§ 5.2). We then ask *where* the gain comes from by decomposing the test-set delta across GREEN error types and tracing that decomposition down to a single SAE feature (§ 5.3). Section 5.4 provides the same pipeline results over two further medical VLMs (LLaVA-RAD and the GRPO-tuned CheXOne). In section 5.5, the resulting boost-suppress feature lists across architectures and provide the structure finding that motivates this paper. We provide qualitative analysis at § 5.6. In Appendix F, we report cross-dataset transfer to IU-Xray as dataset generalization.

5.1 Finding the per-model operating point

We choose causal selection over a Pearson-correlation proxy because it tells us, for each feature, the effect on each GREEN error type sep-

Setting	GREEN (GR↑	RG↑	CXB$_\mu$↑	BS↑	BL↑	RL↑	Comp.↑	RCQ↑
RadVLM	27.7	18.6	47.7	52.9	6.6	25.8	31.5	−1.193
+ SAE Combined (α=0.20, K=20)	**28.8***	**21.0**	**50.3**	**53.0**	**7.0**	25.8	**33.2**	**−1.159**
Δ *pp / abs. (rel.%)*	+1.0 *(+3.9%)*	+2.4 *(+12.9%)*	+2.6 *(+5.5%)*	+0.1 *(+0.2%)*	+0.4 *(+6.1%)*	+0.0 *(+0.2%)*	+1.7 *(+5.4%)*	+0.034 *(–)*
LLaVA-Rad	28.5	16.8	53.3	48.9	5.2	22.2	33.3	−1.374
+ SAE Combined (α=0.50, K=10)	**31.4***	**19.4**	52.6	**54.0**	**6.3**	**24.5**	**35.7**	**−1.276**
Δ *pp / abs. (rel.%)*	+2.9 *(+10.2%)*	+2.6 *(+15.4%)*	−0.7 *(−1.3%)*	+5.1 *(+10.5%)*	+1.1 *(+20.8%)*	+2.3 *(+10.3%)*	+2.4 *(+7.2%)*	+0.098 *(–)*
CheXOne	22.3	18.9	35.5	49.5	4.6	22.7	25.0	−1.339
+ SAE Combined (α=0.50, K=100)	**24.9***	**20.2**	**38.7**	**54.6**	**5.1**	**23.6**	**29.2**	**−1.237**
Δ *(pp / rel.)*	+2.6 *(+11.7%)*	+1.3 *(+6.8%)*	+3.2 *(+9.1%)*	+5.1 *(+10.3%)*	+0.5 *(+10.3%)*	+0.9 *(+4.2%)*	+4.2 *(+17.0%)*	+0.102 *(–)*

Table 1: Steered models evaluation results on the MIMIC-CXR test set. Δ rows show absolute change in percentage points (pp), It uses full–percision RadEval means before rounding. The $rel.\%$ is relative gain w.r.t. the baseline ((SAE-Base)/Base × 100). **RG**: RadGraph F1-simple, **CXB$_\mu$**: CheXbert-14 micro F1, **BS**: BERTScore F1, **BL** and **RL**: BLEU-4 and ROUGE-L, **Comp.**: Composite, **RCQ:** RadCliQ-v1 quality on its native learned scale (higher is better). $*$: $p<0.001$ paired bootstrap on GREEN (#resample=10,000).

arately ($\Delta_{\text{FF}}, \Delta_{\text{MF}}, \Delta_{\text{WL}}, \Delta_{\text{WS}}$), rather than just an overall association with hallucination.

Grid Search. We keep the causal feature ranking and vary and identify the best SAE steering approach on models by varying one design factor at a time on a stratified subset of the validation split ($n = 200$) scored with GREEN, RadGraph, CheXbert-14, and BERTScore. We find that for the steering mode, *residual steering* outperforms its counterpart (SAE-mediated activation patch). Furthermore, the *Combined* approach with $\beta = 1$ gains on every metric on every backbone; suppress-only and boost-only each regress at least one metric on at least one backbone. Among all the α and K combinations, the optimal points are for RadVLM ($\alpha = 0.2$, $K = 20$), for LLaVA-Rad ($\alpha = 0.5$, $K = 10$) and for CheXOne ($\alpha = 0.5$, $K = 100$). We also observe that the optimal α is small and well-defined (0.20–0.50); increasing it breaks coherent reporting and degrades the models' output distribution. On the same subset, steering at four layers $\{8, 16, 20, 24\}$ beats a single layer and reduced subsets.

5.2 RadVLM Results

We evaluate on the held-out MIMIC-CXR test split once with the validation-frozen configuration; Table 1 reports the metrics of that evaluation.

SAE steering improves *every* clinical and text-quality metric on RadVLM (Table 1, top block): GREEN 27.7 → 28.8% (+1.0 pp absolute, +3.9% rel.), RadGraph 18.6 → 21.0% (+12.9% rel.), CheXbert 47.7 → 50.3% (+5.5% rel.), BERTScore 52.9 → 53.0%, RadCliQ −1.193 → −1.159. The Composit gains +1.7 pp absolute, +5.4% relative. NLG metrics also nudge up (BLEU-4 6.6 → 7.0; ROUGE-L a bit more than 25.8). We perform a statistical significance test with paired bootstrap over per-sample scores, resampling $10,000$ from the test set. The GREEN delta is significant at $p < 0.001$.

GREEN error type	Baseline	+ SAE Combined	Δ
False finding (FF)	5,305	6,417	+1112
Missing finding (MF)	7,888	7,234	−654
Wrong location (WL)	1,040	1,208	+168
Wrong severity (WS)	881	953	+72
False comparison (FC)	194	136	−58
Missing comparison (MC)	71	66	−5
Total significant errors	**15,379**	**16,014**	+635

Table 2: Per-error-type breakdown of GREEN *significant* errors for RadVLM on the MIMIC-CXR test set, at the per-token SAE operating point of Table 1.

Error type	#supp.	#boost	Top supp.	Top boost
False finding	100	368	#31611	#15810
Missing finding	68	397	#2541	#1925
Wrong location	155	273	#4240	#28456
Wrong severity	450	14	#29962	#7046

Table 3: Per-error-type causal feature decomposition at $\ell = 16$ (RadVLM, 500 prefiltered features). The count of features whose single-feature ablation reduced (*#supp*) or increased (*#boost*) that error, and the strongest individual feature in each direction.

5.3 Where does the gain come from? Per-error-type decomposition

The composite gain on RadVLM, tells us that the mean report quality improves; it does not say *which* errors are removed. We first decompose the test-set delta across the six GREEN categories (Table 2), then into individual SAE features (Table 3).

The gain on GREEN is an error redistribution, not a free lunch. Steering removes 654 MF errors but adds +1,112 FF, +168 WL, and +72 WS errors (Table 2). GREEN nevertheless improves

because it is structurally *asymmetric in MF*: every recovered missing finding simultaneously adds 1 to the numerator and subtracts 1 from the denominator, whereas a new FF/WL/WS error only enlarges the denominator. The net effect is a **shift toward more complete reporting**: the steered model recovers reference findings that the baseline omits (MF ↓, mean GREEN ↑), occasionally at the cost of mentioning content the reference does not. RadGraph F1, CheXbert-14, BERTScore, and RadCliQ all move in the same direction (Table 1), confirming the gain is not metric gaming. We provide an extended analysis of this trade-off in Appendix C.

Single features carry type-specific causal roles. Table 3 shows that the same SAE coordinate can be a top suppress for one GREEN type and top boost for another (e.g., #1925 is MF-boost and WL-suppress at ℓ=16; #30874 is MF-suppress and FF-boost at ℓ=20). Across the 500-feature pool at ℓ=16, 80 coordinates satisfy $\Delta_{\text{FF}} < 0$ and $\Delta_{\text{MF}} > 0$ simultaneously, justifying separate per-error-type lists in Eq. 2. The per-layer balance also differs: ℓ=24 is almost entirely FF-boost (494/500), while wrong-severity shifts from mixed at ℓ=8 to strongly WS-suppress at ℓ=16 (450 vs. 14). This heterogeneity motivates steering at all four layers (detailed feature activation analysis in Appendix C).

5.4 Cross-architecture generalization

We apply the same pipeline to LLaVA-Rad and CheXOne, training one SAE stack and one feature list per backbone and a validation-selected (α, K). Table 1 shows the results. The composite improves by $+7.2\%$ on LLaVA-Rad and $+17.0\%$ on CheXOne (relative to each model's own baseline). The largest gain is on CheXOne, which is already GRPO-fine-tuned (Zhang et al., 2026b), so one might expect little room for improvement for an inference-time edit; we observe the opposite. SAE residual steering is therefore complementary to GRPO, not redundant with it.

To verify that the identified features belong to the model rather than artifacts of MIMIC-CXR, we apply the exact SAE weights, feature lists and hyperparameters using RadVLM-steered on IU-Xray. We observe a similar trend as GREEN +3.8 pp (+7.7% rel.), RadGraph +4.1 pp (+16.7% rel.), Composite +2.5 pp (+5.4% rel.). In terms of per-error-type pattern, MF decreased by 850 and total errors dropped by 713. Full metrics, per error breakdown, are in Appendix F.

Model pair	Direction	W.Jaccard↑	Cosine↑
RadVLM vs. LLaVA-Rad	Suppress	0.42 [0.33, 0.50]	0.52 [0.36, 0.72]
RadVLM vs. CheXOne	Suppress	0.58 [0.45, 0.71]	0.46 [0.24, 0.67]
LLaVA-Rad vs. CheXOne	Suppress	0.41 [0.25, 0.62]	0.25 [0.12, 0.40]
RadVLM vs. LLaVA-Rad	Boost	0.65 [0.60, 0.75]	0.93 [0.91, 0.96]
RadVLM vs. CheXOne	Boost	0.75 [0.56, 0.94]	0.95 [0.94, 0.95]
LLaVA-Rad vs. CheXOne	Boost	0.75 [0.65, 0.80]	0.91 [0.88, 0.94]

Table 4: Cross-model functional overlap with 95% bootstrap Confidence intervals (CIs) (mean [lo, hi]). Per-cell aggregate is the mean over steered layers $\{8, 16, 20, 24\}$ of the top-100 features per direction; CIs are percentile-bootstrap intervals from $10,000$ resamples of the four layer values with replacement, capturing cross-layer stability. Boost CIs are tight and lie well above suppress CIs in every model pair.

5.5 Cross-Model Feature Alignment: What Generalizes and What Does Not

Selecting features separately for each model raises a mechanistic question: *how many of the underlying features are actually shared across models?* Since SAE feature indices are not directly comparable across architectures, we shift the analysis to a functional perspective. Concretely, for each pair of models, we compute the weighted Jaccard (Ruzicka) similarity between the GREEN error-type profiles of the top 100 features, stratified by direction and layer (see Appendix E). In other words, we check how similar two models are by looking at their most important features and asking: do they affect the same types of errors? We measure this overlap using weighted Jaccard similarity, based on which error types each feature mainly influences, as measured during feature identification (§3). Across all model pairs, the boost features are quite consistent: about 65–75% of them overlap, and their overall behavior is very similar (cosine around 0.91–0.95, with 95% confidence intervals of about ± 0.03. It indicates that features that improve quality tend to work in the same way across different models. In contrast, the suppress features are much less consistent. Their overlap is lower (41–58%), and in some cases their behavior differs a lot. For example, between LLaVA-Rad and CheXOne, the similarity drops sharply. This suggests that each model handles hallucinations in its own way, relying on different internal mechanisms. Importantly, this difference is very clear statistically: for every model pair, the similarity of boost features is always higher than the similarity of suppress

features, with no overlap between their confidence ranges. So this pattern is robust, and it's not just due to picking specific layers.

5.6 Qualitative examples

Due to MIMIC-CXR's data-use agreement, we reproduce RadVLM steering results on IU-Xray. Table 15 (Appendix F) gives four pathological cases with GREEN tagged baseline vs. steered reports.

6 Discussion and Conclusion

The GREEN improvements we report on three radiology VLMs, are statistically significant but moderate in absolute terms compared with what task-specific post-training or RL delivers. We therefore do not present the method as a replacement for post-tuning. Instead, the main contribution of our approach is providing a way to introspect and open parts of the model's black box. With per-layer SAEs and causal feature screening, as *a mechanistic window* into a medical VLM's residual stream, a medical VLM can be described in terms of a few hundred sparse coordinates whose effects can be named, located in the report, and edited at inference time. This approach avoids any weight updates and substantially reduces GPU-hour costs per model compared with RL. The GREEN gains show that these features are not only interpretable; they are also faithful enough to steer.

Hallucination has type-specific structure, not a single "hallucination direction". The feature analysis shows that hallucination is not one global direction that can simply be turned down. The same SAE coordinate can help one GREEN error type while hurting another: for example, #1925 appears in MF-boost and WL-suppress, and #30874 appears in MF-suppress and FF-boost (Table 3). At ℓ=16, $80/500$ screened coordinates simultaneously satisfy $\Delta_{\mathrm{FF}} < 0$ and $\Delta_{\mathrm{MF}} > 0$. This is why we keep separate suppress and boost lists for each error type. The structure also changes across layers: ℓ=24 is almost entirely FF-boost ($494/500$), while wrong-severity features move from mixed at ℓ=8 to strongly WS-suppress at ℓ=16 (450 vs. 14). This kind of per-error, per-layer organization is hidden by aggregate scores.

Trade-offs are localizable and transparent. The main trade-off is also interpretable. Combined steering recovers 654 missing findings but adds $1,112$ false findings (Table 2), which could look like a noisier model. The SAE features show a more specific story. One sparse, late-report feature (#15810; $43/9,951$ studies, 87% in the late half) is activated on repeated phrases, and two dense features (#5872, #16965) fire on generic normal-report templates. On the other side, features #2541 and #4240 activate on real findings and end-of-report negation lists (Tables 11, 12). The trade-off is therefore not random fabrication; it is more complete reporting with some templated repetition. This also explains why GREEN, RadGraph, CheXbert, and RadCliQ improve together despite the FF increase: recovering a missing finding has a larger effect under GREEN's structure than adding one denominator-only error (Eq. 5). Because the edit is controlled by α and by per-error-type lists, a user who needs a stricter FF limit can adjust α on a local validation set.

What transfers across models, and what does not. The cross-model alignment study shows a simple message. Boost features look similar across architectures (signature cosine 0.91-0.95), but suppress features are much more model-specific (as low as 0.25 $[0.12, 0.40]$ for LLaVA-Rad vs. CheXOne; Table 4). In other words, models share some quality-promoting directions, but each model develops its own hallucination pathways. This matters for transfer: boost priors may transfer, while suppress features should be re-examined for each backbone. The largest gain appears on CheXOne, even though it is already GRPO-tuned, suggesting that SAE steering and RL are complementary rather than redundant.

As an **outlook**, our mechanistic study opens concrete follow-up work that aggregate metric evaluation alone does not suggest: census-aligned steering recipes that share boost priors across models while specialising suppress weights per backbone; additional inference-time baselines (ITI (Li et al., 2023), DoLa(Chuang et al., 2024)) to compare with other lightweight editing methods; joint language-stream and vision-stream steering (Pach et al., 2026); and extension beyond chest radiography. More broadly, we hope that demonstrating *interpretable* hallucination mitigation, where every edit maps to a named SAE feature, a measurable causal effect, and a localisable position in the generated report, encourages the field to prioritize mechanistic transparency alongside advances in evaluation metrics.

Limitations

Our study is offline, on MIMIC-CXR and IU-Xray English reports; clinical deployment requires regulatory scrutiny well beyond the scope of this paper. GREEN and the other automatic metrics we use are proxies for radiologist judgment; the GREEN error taxonomy may also miss important clinical distinctions. We exclude comparison-type errors (false/missing comparison) because the subset of MIMIC-CXR that we employed does not include prior studies for comparison. Generalization beyond chest radiography is an open question. Because of MIMIC's data usage agreement, we provide our quantitative examples using the IU-Xray dataset. Table 1 applies the *same* SAE steering protocol to *RadVLM, LLaVA-Rad, and CheXOne* (three architectures: two supervised, one GRPO-tuned), each with its own SAE and a one-shot test evaluation; further backbones (e.g., other Qwen-family models) and census-driven steering variants beyond the validated Combined recipe are left to future work. The SAE dictionary size and sparsity introduce hyperparameters whose interactions with model scale are not fully characterized. Finally, we intervene on the LLM's residual stream only and not on the vision encoder. Pach et al. (2026) shows that vision-stream SAEs can also steer multimodal output without language-model edits. A natural extension is to combine the two intervention sites for medical VLMs, using language-stream features to suppress hallucinated phrasing (this work) and vision-stream features to amplify or attenuate radiographic concepts at an earlier stage, and we leave that combination to future work.

Ethics Statement

This work uses the publicly available MIMIC-CXR-JPG dataset under its established data-use agreement and the IU-Xray dataset. The method is intended for research and is not validated for clinical decision-making. Automated report-generation systems, steered or unsteered ones, should not replace radiologist review in clinical practice. Any deployment would require prospective validation and regulatory approval.

We used an LLM assistant for limited rewriting and paraphrasing of parts of the paper; all experiments, analyses, and final claims in this manuscript were produced and verified by the authors.

References

Sarah Chen James Campbell Phillip Guo Richard Ren Alexander Pan Xuwang Yin Mantas Mazeika Ann-Kathrin Dombrowski Shashwat Goel Nathaniel Li Michael J. Byun Zifan Wang Alex Mallen Steven Basart Sanmi Koyejo Dawn Song Matt Fredrikson Zico Kolter Dan Hendrycks Andy Zou, Long Phan. 2023. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

Dana Arad, Aaron Mueller, and Yonatan Belinkov. 2025. SAEs are good for steering – if you select the right features. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10241–10259, Suzhou, China. Association for Computational Linguistics.

Shruthi Bannur, Kenza Bouzid, Daniel Coelho de Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Fabian Falck, Ozan Oktay, Anja Thieme, Matthew P Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie Hyland. 2024. Maira-2: Grounded radiology report generation. Technical Report MSR-TR-2024-18, Microsoft.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, and 1 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2(5):6.

Amos Calamida, Farhad Nooralahzadeh, Morteza Rohanian, Koji Fujimoto, Mizuho Nishio, and Michael Krauthammer. 2023. Radiology-aware model-based evaluation metric for report generation. *arXiv preprint arXiv:2311.16764*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, volume 2024, pages 54158–54183.

Dina Demner-Fushman and 1 others. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Nicolas Deperrois, Hidetoshi Matsuo, Samuel Ruiperez-Campillo, Moritz Vandenhirtz, Sonia Laguna, Alain Ryser, Koji Fujimoto, Mizuho Nishio, Thomas Sutter, Julia Vogt, Jonas Kluckert, Thomas Frauenfelder, Christian Bluethgen, Farhad Nooralahzadeh, and Michael Krauthammer. 2025a. RadVLM Instruction Dataset. *PhysioNet*. Version 1.0.0.

Nicolas Deperrois, Hidetoshi Matsuo, Samuel Ruipérez-Campillo, Moritz Vandenhirtz, Sonia Laguna, Alain Ryser, Koji Fujimoto, Mizuho Nishio, Thomas M.

Sutter, Julia E. Vogt, Jonas Kluckert, Thomas Frauenfelder, Christian Blüthgen, Farhad Nooralahzadeh, and Michael Krauthammer. 2025b. Radvlm: A multitask conversational vision-language model for radiology. *ArXiv*, abs/2502.03333.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2025. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*.

Benjamin Gundersen, Nicolas Deperrois, Samuel Ruiperez-Campillo, Thomas M. Sutter, Julia E Vogt, Michael Moor, Farhad Nooralahzadeh, and Michael Krauthammer. 2026. RadVLM-GRPO: Enhancing chest x-ray report generation and visual grounding via reinforcement learning. In *Medical Imaging with Deep Learning*.

Dennis Hein, Zhihong Chen, Sophie Ostmeier, Justin Xu, Maya Varma, Eduardo Pontes Reis, Arne Edward Michalson, Christian Bluethgen, Hyun Joo Shin, Curtis Langlotz, and Akshay S Chaudhari. 2025. CheXalign: Preference fine-tuning in chest X-ray interpretation models without human feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27679–27702, Vienna, Austria. Association for Computational Linguistics.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis Langlotz, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. 2024. MIMIC-CXR Database. *PhysioNet*. Version 2.1.0.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*.

Kyle O'Brien, David Majercak, Xavier Fernandes, Richard G. Edgar, Blake Bullwinkel, Jingya Chen, Harsha Nori, Dean Carignan, Eric Horvitz, and Forough Poursabzi-Sangdeh. 2025. Steering language model refusal with sparse autoencoders. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. 2024. GREEN: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 374–390, Miami, Florida, USA. Association for Computational Linguistics.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. 2026. Sparse autoencoders learn monosemantic features in vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Joris Postmus and Steven Abreu. 2024. Steering large language models using conceptors: Improving addition-based activation engineering. In *MINT: Foundation Model Interventions*.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online. Association for Computational Linguistics.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and explainable region-guided radiology report generation. In *CVPR*.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, and 3 others. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Justin Xu, Xi Zhang, Javid Abderezaei, Julie Bauml, Roger Boodoo, Fatemeh Haghighi, Ali Ganjizadeh, Eric Brattain, Dave Van Veen, Zaiqiao Meng, David W Eyre, and Jean-Benoit Delbrouck. 2025. RadEval: A framework for radiology text evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 546–557, Suzhou, China. Association for Computational Linguistics.

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. 2022. Evaluating progress in automatic chest x-ray radiology report generation. *medRxiv*.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, and 8 others. 2025. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yabin Zhang, Chong Wang, Yunhe Gao, Jiaming Liu, Maya Varma, Justin Xu, Sophie Ostmeier, Jin Long, Sergios Gatidis, Seena Dehkharghani, Arne Michalson, Eun Kyoung Hong, Christian Bluethgen, Haiwei Henry Guo, Alexander Victor Ortiz, Stephan Altmayer, Sandhya Bodapati, Joseph David Janizek, Ken Chang, and 3 others. 2026a. A reasoning-enabled vision-language foundation model for chest x-ray interpretation. *Preprint*, arXiv:2604.00493.

Yabin Zhang, Chong Wang, Yunhe Gao, Jiaming Liu, Maya Varma, Justin Xu, Sophie Ostmeier, Jin Long, Sergios Gatidis, Seena Dehkharghani, Arne Michalson, Eun Kyoung Hong, Christian Bluethgen, Haiwei Henry Guo, Alexander Victor Ortiz, Stephan Altmayer, Sandhya Bodapati, Joseph David Janizek, Ken Chang, and 3 others. 2026b. A reasoning-enabled vision-language foundation model for chest x-ray interpretation. *Preprint*, arXiv:2604.00493.

Appendix

A Layer selection

We train SAEs at ten evenly-spaced depths $\{0, 4, 8, 12, 16, 20, 24, 28, 32, 35\}$ of RadVLM's 36-layer Qwen3-VL backbone but hook only a subset $\mathcal{L}=\{8, 16, 20, 24\}$ for steering. Table 5 reports SAE reconstruction quality at each depth, revealing three regimes that motivate this choice.

Early layers ($\ell \leq 4$): poor decomposition. The SAE struggles to reconstruct these activations faithfully. Early transformer layers mainly encode positional and syntactic information rather than clinical content (Templeton et al., 2024), so there is little to steer.

Late layers ($\ell \geq 28$) : norms too large. Reconstruction quality is fine, but activation norms explode (244–1,412 vs. 31–128 in the steered range). Our steering perturbation δ is a fixed-size nudge; when the residual stream is 10–45$\times$ larger, that nudge becomes negligible unless α is cranked up to values that break generation. These layers are also close to the output and mainly encode next-token predictions, not factual content (Gao et al., 2025; Bricken et al., 2023).

Middle-to-late layers ($8 \leq \ell \leq 24$) : the sweet spot. Cosine similarity is high (> 0.97) and norms stay moderate, so the SAE can both decompose and meaningfully edit the hidden states. We pick $\{8, 16, 20, 24\}$ as roughly quarter-spaced depths through this range, consistent with where prior work finds semantic and factual features in LLMs (Templeton et al., 2024; Marks et al., 2025). Layer 12 has comparable quality and could be added; we leave it out to keep the causal screening budget tractable ($\sim$5 extra GPU-h per layer). We observe that a multi-layer ablation in §5.1 confirms that steering all four layers outperforms any single-layer or reduced subset.

B Compute

Table 6 compares the compute cost of SAE steering against GRPO fine-tuning. The full SAE pipeline, activation collection through causal feature identification, takes roughly 27 GPU-hours on NVIDIA GH200 for a single model (RadVLM); SAE training itself is < 1 GPU-hour. GRPO fine-tuning takes multiple days of RL training per model. At inference time, SAE steering is essentially free: one SAE encode/decode per layer per token, on tensors

Layer	Cos. sim. ↑	Dead % ↓	$\|\mathbf{h}\|$
0	0.993	63.9	8
4	0.992	78.0	22
8	0.985	77.3	31
12	0.981	81.7	38
16	0.982	81.9	49
20	0.974	82.6	64
24	0.969	84.6	128
28	0.970	83.1	244
32	0.980	82.5	510
35	0.994	82.4	1412

Table 5: **SAE reconstruction quality across all ten trained layers** (RadVLM, D=32,768, k=64). *Cos. sim.*: mean cosine similarity between input and reconstruction (> 0.97 for all steered layers). *Dead %*: fraction of dictionary features never activated. $\|\mathbf{h}\|$: mean activation norm. **Bold** layer numbers mark the steered set $\mathcal{L}=\{8, 16, 20, 24\}$. Horizontal rules separate three regimes: early layers (0–4) with poor feature decomposition; middle-to-late layers (8–24) with high reconstruction quality and moderate norms; late layers (28–35) where norms grow 8–45×, making steering perturbations relatively small.

Component	SAE steering	GRPO
Requires training?	No (frozen weights)	Yes
Activation collection	∼6 GPU-h	—
SAE training (10 layers)	<1 GPU-h	—
Causal feature identification	∼20 GPU-h	—
RL training	—	∼350 GPU-h
Total pipeline	**∼27 GPU-h**	∼350 **GPU-h**
Inference overhead	<5%	0%
Per-model retraining needed?	Yes (SAEs)	Yes (full RL)
Interpretable features?	✓	✗
Per-error-type control?	✓	✗

Table 6: Compute and capability comparison between SAE steering and GRPO fine-tuning. GPU-hours for SAE work are on NVIDIA GH200 (96GB). The bulk of SAE steering's cost is the causal feature-identification stage (single-feature forward screens, ∼27 GPU-h total across all four steered layers, sharded over 4 GPUs). The GRPO column is the RL training cost reported by RadVLM-GRPO (Gundersen et al., 2026)

.

already in GPU memory (<5% wall-clock overhead).

C Extended error-redistribution analysis

This section provides the extended discussion of the error-redistribution trade-off summarised in §5.3.

GREEN's structural asymmetry. Let M be the number of matched findings in a report and E_i the count of significant errors of type $i \in \{$(a) FF, (b) MF, (c) WL, (d) WS, (e) false comparison, (f) missing comparison$\}$. Then

$$\text{GREEN} = \frac{M}{M + \sum_{i=(\text{a})}^{(\text{f})} E_i} \in [0, 1], \quad (5)$$

defined as 0 when M=0. A single MF→match conversion therefore moves GREEN by roughly twice as much as one denominator-only addition (adding 1 to numerator *and* subtracting 1 from denominator), so the score can rise even when the total error count grows. In our case 654 MF errors flip to matches while the additional +1,352 FF/WL/WS errors land on the denominator-only side, and the ratio still increases. RadGraph entity-relation F1 (recall-weighted) and RadCliQ (calibrated to radiologist preference, smaller penalty on added than on missed content) carry the same asymmetry, which is why all three metrics move together at this operating point.

Why the total error rise is not a uniform quality loss. Two facts are easy to overlook. **First**, mean GREEN improves, and RadGraph F1, CheXbert-14, BERTScore, and RadCliQ all move in the same direction (Table 1). A model trading missing findings for unrelated false findings would not improve RadGraph or CheXbert similarity. **Second**, GREEN's six categories are clinically asymmetric: a missed *pneumothorax* is more harmful than a redundantly reported "normal-variant cardiac silhouette", and per-report mean GREEN, RadGraph, and CheXbert weight recall of reference findings more than added content.

What do the boost features actually activate on? We re-encode the RadVLM per-token validation activation pool (10,000 decodes, 9,951 distinct study_ids) through the SAE at layer 16 and aggregate token-level pre-activations. The FF-boost set separates into two categories: (i) a sparse feature associated with late-stage *repetition loops* (#15810; 43/9,951 studies, 87% in the latter half, top activations on direct phrase repetition); and (ii) dense template features (#5872, #16965) that activate across nearly all studies and reflect *generic report-normal* structures. In contrast, the MF/WL suppress examples #2541 and #4240 peak on paired-finding wording and end-of-report negation chains (Table 11); ablating #2541 lowers MF (Δ_{MF}=−0.09) while raising FF (Δ_{FF}=+0.14), and ablating #4240 lowers WL (Δ_{WL}=−0.09). Together, these observations suggest that the +1,112 FF increase is better interpreted as a consequence of more complete late-report generation, rather

Error Type	Role	Feat.	Δ_{FF}	Δ_{MF}	Δ_{WL}	Δ_{WS}
False finding	supp.	6596	-0.310	$+0.040$	-0.040	-0.010
		24597	-0.260	-0.040	$+0.010$	$+0.030$
		94	-0.240	-0.020	$+0.040$	$+0.000$
	boost	18416	$+0.200$	$+0.060$	-0.040	-0.020
		12906	$+0.190$	$+0.060$	-0.030	-0.020
		11500	$+0.180$	$+0.040$	-0.030	-0.010
Missing finding	supp.	3983	-0.180	-0.130	$+0.020$	$+0.040$
		23671	-0.120	-0.120	$+0.010$	$+0.030$
		18669	-0.140	-0.110	$+0.010$	$+0.020$
	boost	23148	$+0.090$	$+0.370$	-0.100	-0.100
		1542	$+0.150$	$+0.250$	-0.100	-0.070
		26996	-0.110	$+0.240$	-0.020	$+0.020$
Wrong location	supp.	13589	$+0.070$	$+0.200$	-0.130	-0.040
		1542	$+0.150$	$+0.250$	-0.100	-0.070
		6658	$+0.000$	$+0.150$	-0.100	-0.030
	boost	17609	-0.100	-0.010	$+0.070$	$+0.040$
		19414	-0.070	-0.020	$+0.060$	-0.060
		19133	-0.050	-0.020	$+0.050$	$+0.000$
Wrong severity	supp.	23148	$+0.090$	$+0.370$	-0.100	-0.100
		21739	$+0.010$	$+0.100$	-0.060	-0.080
		31838	$+0.050$	-0.030	-0.010	-0.080
	boost	7210	-0.110	-0.020	$+0.020$	$+0.070$
		26264	$+0.020$	$+0.030$	$+0.000$	$+0.060$
		6916	-0.080	-0.090	$+0.050$	$+0.060$

Table 7: **Top causal features at layer 8.** Top-3 suppress and boost features for each GREEN error type, with deltas on FF/MF/WL/WS.

Error Type	Role	Feat.	Δ_{FF}	Δ_{MF}	Δ_{WL}	Δ_{WS}
False finding	supp.	31611	-0.160	$+0.040$	-0.030	-0.040
		23239	-0.130	-0.030	-0.030	-0.030
		8320	-0.130	$+0.100$	-0.070	-0.050
	boost	15810	$+0.300$	$+0.080$	$+0.000$	-0.020
		5872	$+0.280$	$+0.080$	$+0.000$	-0.100
		16965	$+0.270$	$+0.060$	-0.060	-0.060
Missing finding	supp.	2541	$+0.140$	-0.090	-0.020	-0.040
		25211	$+0.020$	-0.070	-0.010	$+0.020$
		8457	$+0.130$	-0.060	-0.020	$+0.000$
	boost	1925	$+0.150$	$+0.370$	-0.080	$+0.000$
		20872	$+0.110$	$+0.230$	-0.020	-0.080
		9114	$+0.050$	$+0.190$	-0.020	-0.040
Wrong location	supp.	4240	$+0.200$	$+0.060$	-0.090	-0.030
		22773	$+0.010$	$+0.140$	-0.080	-0.070
		1925	$+0.150$	$+0.370$	-0.080	$+0.000$
	boost	28456	-0.030	$+0.040$	$+0.060$	-0.020
		16975	$+0.070$	$+0.000$	$+0.060$	-0.030
		11553	$+0.080$	-0.020	$+0.050$	-0.010
Wrong severity	supp.	29962	$+0.180$	$+0.080$	$+0.020$	-0.110
		13311	$+0.160$	$+0.060$	-0.010	-0.110
		8368	$+0.220$	$+0.170$	-0.030	-0.100
	boost	7046	$+0.040$	$+0.010$	-0.040	$+0.050$
		24384	$+0.060$	$+0.090$	$+0.020$	$+0.030$
		31685	$+0.110$	$+0.040$	$+0.010$	$+0.020$

Table 8: **Top causal features at layer 16.** Top-3 suppress and boost features for each GREEN error type, with deltas on FF/MF/WL/WS.

than evidence of safety-critical fabrication. *To our knowledge, this is the first quantitative localisation of SAE features in a medical VLM.*

Layer-wise heterogeneity details. Among the screened features at ℓ=24, almost every coordinate is FF-boost signed ($494/500$ with $\Delta_{\text{FF}} > 0$), whereas wrong-severity at ℓ=8 still mixes WS-suppress and WS-boost coordinates (234 vs. 184), but at ℓ=16 this distribution becomes highly skewed toward suppression (450 vs. 14). Our ablation results further support the multi-layer design: removing any individual layer from $\mathcal{L}=\{8, 16, 20, 24\}$ leads to a measurable degradation rather than being compensated by the remaining layers.

D Interpretation: Features, behaviour, and cases

Here, we walk through what the screened SAE features do in practice, in four progressively high-level subsections. We first list the strongest causal features per steered layer and the text contexts they activate on (Tables 7-10 and Table 11); we then quantify *where* in the report each reported features activates (Table 12, Figure 4), we visualize the full distribution of per-error-type causal effects $\Delta_j^{(j)}$ at the canocnical layer $\ell = 16$ in Figure 3. A short synthesis paragraph at the end relates these observations to the cross-model boost-suppress finding in § 5.5.

Top causal features per steered layer. Tables 7-10 list, for each steered layer and each GREEN error type, top-3 suppress (most negative per-type delta) and top-3 boost (most positive per-type delta) features among the 500 prefiltered candidates from the causal selection pipeline. The same SAE feature can appear in multiple of these eight (error-type $\times$ role) sub-blocks at the same time. For example, feature #1925 at layer 16 sits in the MF-*boost* block ($\Delta_{\text{MF}} = +0.37$) and in the WL-*suppress* block ($\Delta_{\text{WL}} = -0.08$): the same coordinate plays opposite causal roles for different clinical mistakes.

Figure 3 visualises the full distribution of per-type deltas at ℓ=16. The figure has *four panels*, one per GREEN error type (FF / MF / WL / WS); each panel is a histogram of one number per feature.

For every feature j in the 500-feature pool at ℓ=16 we measured its causal effect $\Delta_t^{(j)}$ on a *single* error type t, the average change in that error's count per report when j is zeroed on the N=100 screening slice (§3). That is one number per feature per error type, so the panel for type t contains exactly 500 values of $\Delta_t^{(j)}$.

The *x-axis* is that value $\Delta_t^{(j)}$ (mean errors per report). The *y-axis* is the number of features whose Δ falls in that x-bin as a standard histogram count, not a clinical metric. The interpretation of the bars with different colors is as follows:

- **red bars (left of 0):** features for which zeroing *reduces* that error type , the feature was *promoting* the error, so it is a *suppress* candidate for that

Error type	Role	Feat. #	Δ_{FF}	Δ_{MF}	Δ_{WL}	Δ_{WS}
False finding	supp.	20624	-0.220	-0.020	$+0.020$	$+0.010$
		21539	-0.210	-0.090	$+0.000$	$+0.030$
		22189	-0.210	-0.020	$+0.010$	$+0.000$
	boost	30874	$+0.330$	-0.140	$+0.030$	$+0.090$
		14416	$+0.270$	$+0.060$	-0.020	$+0.080$
		17767	$+0.260$	$+0.030$	-0.010	-0.010
Missing finding	supp.	30623	$+0.000$	-0.150	$+0.010$	$+0.050$
		30874	$+0.330$	-0.140	$+0.030$	$+0.090$
		16748	-0.100	-0.140	$+0.000$	$+0.040$
	boost	24418	$+0.050$	$+0.220$	-0.050	-0.050
		16881	$+0.090$	$+0.170$	-0.010	-0.030
		5385	$+0.100$	$+0.150$	-0.030	-0.020
Wrong location	supp.	9659	-0.090	$+0.030$	-0.110	-0.010
		15819	-0.070	$+0.080$	-0.080	-0.010
		2651	$+0.010$	-0.050	-0.080	$+0.020$
	boost	5091	-0.110	-0.130	$+0.050$	$+0.020$
		26348	-0.140	-0.110	$+0.050$	$+0.030$
		29751	-0.060	-0.090	$+0.050$	$+0.030$
Wrong severity	supp.	5571	-0.180	$+0.050$	-0.010	-0.060
		24418	$+0.050$	$+0.220$	-0.050	-0.050
		22955	-0.040	$+0.070$	$+0.000$	-0.040
	boost	30874	$+0.330$	-0.140	$+0.030$	$+0.090$
		14416	$+0.270$	$+0.060$	-0.020	$+0.080$
		3685	-0.100	-0.060	$+0.020$	$+0.080$

Table 9: **Top causal features at layer 20.** Top-3 suppress and boost features for each GREEN error type, with deltas on FF/MF/WL/WS.

Error type	Role	Feat. #	Δ_{FF}	Δ_{MF}	Δ_{WL}	Δ_{WS}
False finding	supp.	14217	-0.020	$+0.110$	-0.020	-0.010
		1685	-0.020	$+0.000$	-0.010	-0.030
		10884	-0.020	$+0.000$	-0.010	-0.040
	boost	20420	$+0.560$	$+0.010$	-0.070	$+0.090$
		3026	$+0.360$	$+0.040$	-0.030	-0.010
		15397	$+0.310$	-0.030	-0.050	$+0.010$
Missing finding	supp.	1972	$+0.110$	-0.140	-0.040	$+0.010$
		24480	$+0.200$	-0.080	-0.030	-0.050
		20968	$+0.060$	-0.070	-0.010	$+0.030$
	boost	15720	$+0.080$	$+0.270$	-0.020	-0.090
		16413	$+0.200$	$+0.150$	-0.040	-0.050
		15291	$+0.020$	$+0.130$	-0.010	$+0.030$
Wrong location	supp.	28349	$+0.130$	$+0.080$	-0.110	-0.010
		32155	$+0.180$	$+0.120$	-0.100	-0.060
		25435	$+0.210$	$+0.050$	-0.100	-0.110
	boost	22492	$+0.190$	$+0.010$	$+0.020$	$+0.000$
		11120	$+0.120$	$+0.010$	$+0.020$	$+0.010$
		25793	$+0.160$	-0.020	$+0.020$	$+0.010$
Wrong severity	supp.	25435	$+0.210$	$+0.050$	-0.100	-0.110
		15720	$+0.080$	$+0.270$	-0.020	-0.090
		12903	$+0.180$	$+0.090$	-0.030	-0.080
	boost	20420	$+0.560$	$+0.010$	-0.070	$+0.090$
		10254	$+0.050$	$+0.010$	-0.030	$+0.040$
		16654	$+0.140$	-0.010	-0.020	$+0.040$

Table 10: **Top causal features at layer 24.** Top-3 suppress and boost features for each GREEN error type, with deltas on FF/MF/WL/WS.

type.

- **green bars (right of 0):** features for which zeroing *increases* that error type, the feature was *preventing* the error, so it is a *boost* candidate for that type.
- **bars at 0:** features whose ablation barely moves this error type (most of the pool, by design, the prefilter only guarantees activation-difference relevance, not per-type causal sign).

The reason that we show in the four panels is that the same feature shows up as a *different* number in each panel, because $\Delta_t^{(j)}$ depends on the error type t. A given feature can sit in the left tail of one panel and in the right tail of another at the same time, e.g., feature #1925 contributes to the MF-boost (right) tail and the WL-suppress (left) tail simultaneously. That per-feature asymmetry is exactly what forces separate suppress and boost lists *per error type* in the combined intervention (Eq. 2).

Features are identified by causal effect, not by meaning. We do *not* hand-label individual SAE features with clinical concepts. A feature's identity throughout this paper is its causal effect on each error type ($\Delta_t^{(j)}$, §3), not a name or a lexical definition. Table 11 is a complementary, *observational* diagnostic: for each reported feature, it shows the unsteered text that activates that feature most strongly.

We generated $10,000$ validation reports with the unsteered RadVLM (*no SAE editing*) and pushed every layer-16 hidden state through the SAE encoder. For each reported feature, we kept the three word positions across that entire pool with the largest SAE activation, and printed a $\sim$100-character window of generated text around each one. This "top-activating contexts" practice for interpreting individual SAE features follows Bricken et al. (2023) and Templeton et al. (2024). To better understand Table 11, here we explain each of its columns: `Feature`: the SAE coordinate index j. `Role/target`: the per-error-type label assigned in §3 (suppress / boost) with that feature's causal $\Delta_t^{(j)}$ on its target error type, repeated here for context (not re-measured in this table). `Act.`: the SAE activation value z_j of feature j at that single word position. Concretely it is one dot product plus one scalar bias,

$$z_j \;=\; \big(h - b_{\text{dec}}\big)^{\top} W_{\text{enc}}[:, j] \;+\; b_{\text{enc},j},$$

where $h \in \mathbb{R}^{4096}$ is the layer-16 residual-stream hidden state at that word and $W_{\text{enc}}[:, j]$ is feature j's learned "fingerprint" direction (the centred-linear encoder of Bricken et al. (2023); we use the TopK variant of Gao et al. (2025)). Bigger z_j means h aligns more strongly with that direction; we bypass the per-token Top-K mask so z_j is interpretable even for features outside that token's top-K. `Generated-report context`: a $\sim$100-character window of the unsteered report around that word, plus the source `study_id`.

It is worth mentioning that the Table 11 is not a baseline-vs-steered comparison. The SAE only *inspects* activations in this table; it does not edit them;

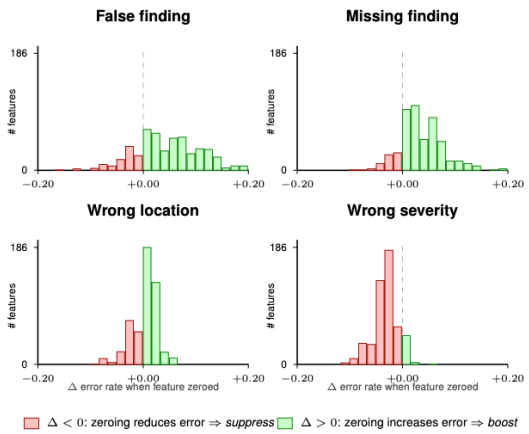


Figure 3: **Distribution of per-error-type causal effects at layer 16 (RadVLM, 500 prefiltered features).** Each panel shows the histogram of $\Delta_{\text{type}}=\text{error}_{\text{ablated}} - \text{error}_{\text{baseline}}$ for one GREEN error type. Left tail (red): suppress candidates (ablation reduces this error). Right tail (green): boost candidates. The asymmetric tails motivate the per-error-type lists used in the combined intervention: features with strong $\Delta_{\text{FF}}<0$ are not the same set as features with strong $\Delta_{\text{MF}}>0$.

every report shown is unsteered RadVLM output. The diagnostic answers "where in the unsteered output does each feature activate?", not "how much did steering improve the report?". The steering result is in Tables 1, and 2.

Each feature has a causal sign that tells us what it does. A *boost* feature has $\Delta_t>0$: turning it off makes error type t worse, so it was helping the model when it was activated. A *suppress* feature has $\Delta_t<0$: turning it off makes the error t better, so it was driving that error when it activated. The text snippets in Table 11 therefore have a simple interpretation: boost firings show what the model looks like when it is doing something right, and suppress activations show what it looks like when it is making mistakes.

From snippets to hypothesized patterns. Inspecting the top-3 firings of each reported feature, we find that the surrounding text clusters into a small number of recurring phrasings, conventions, and fixed templates that the unsteered model produces repeatedly. For each feature, we summarize our interpretation of the dominant theme as a short label, shown in the Hypothesized pattern column of Table 11 as an interpretive annotation aid for the snippets, not a quantity derived from data. Across the 12 reported features, the patterns split cleanly along the boost / suppress axis: *boost* features activate on radiology-writing *habits*, and *suppress* features activate on *template traps*. The two paragraphs below describe each class.

Boost features encode radiology-writing habits. The boost features activate on three habits of careful radiology writing. **Explicit negation**: stating clearly that something is *not* present. #16965 ($\Delta_{\text{FF}}=+0.27$) activates on *"There is no evidence of pneumothorax. There is no evidence of pulmonary edema"*. **Anatomic checklist**: walking through lungs → mediastinum → bones. #5872 ($\Delta_{\text{FF}}=+0.28$) activates on *"The lungs are clear. The cardiomediastinal silhouette is within normal"*; #20872 ($\Delta_{\text{MF}}=+0.23$) continues with *"... the osseous..."*, the cue to enumerate skeletal findings. **Listing incidental findings**: noting old fractures, implants, devices. #1925 ($\Delta_{\text{MF}}=+0.37$, the largest boost effect) activates on *"Multiple old bilateral rib fractures"* and *"A vagal nerve stimulator is present"*. Amplifying these features makes the model more thorough and more grounded in the image; turning them off makes the model less careful about them.

Suppress features mark *template traps*. A template trap is a fixed phrase the model commits to and produces in full once it begins, regardless of what the image actually shows. Three traps appear among the reported suppress features.

End-of-report negation chain. #4240 ($\Delta_{\text{WL}}=-0.09$) activates on *"No pleural effusions. No pulmonary edema. No pneumonia."*. **Paired-finding listing.** #2541 ($\Delta_{\text{MF}}=-0.09$, $\Delta_{\text{FF}}=+0.14$) peaks on effusion/atelectasis-style paired clauses in its top contexts; ablation reduces MF on the screening slice despite that wording, so we treat the causal Δ as authoritative and the snippet as localization only (not a literal map from zeroing the feature to deleting every effusion phrase). **Generic-normal severity template.** #29962 ($\Delta_{\text{WS}}=-0.11$) activates on *"... The aorta is tortuous. There is no pneumothorax or pleural effusion"*. **Confident support-device language.** #31611 ($\Delta_{\text{FF}}=-0.16$) activates on *"lead tip in the distal SVC"* and *"right-sided Swan-Ganz catheter"*. Turning off such a feature does not stop the model from describing real devices or real negative findings; it makes the model less likely to enter the template *by default* when the image does not warrant it.

Opposing roles on the same behaviour. Study s58637592 appears in the top activation contexts of two features with opposite roles. #15810 (boost, $\Delta_{\mathrm{FF}}{=}{+}0.30$) activates on *"... is a prosthetic aortic valve. There is a prosthetic aortic valve.... "*, the start of a repetition loop; keeping it active prevents the loop from cascading into more false findings. #8368 (suppress, $\Delta_{\mathrm{WS}}{=}{-}0.10$) activates on the same loop in the same study, but here it contributes to wrong-severity drift. The two features react to the same underlying behaviour and pull in opposite directions on different error types. This is why the steering recipe needs a separate suppress and boost list *per error type* (§3): a single global label per feature would get one of the two effects wrong (Table 3).

Synthesis and link to the cross-model functional overlap. Previously, we observed that: *boost features amplify radiology-writing habits, suppress features close the model's own template traps, and combined steering applies both.* This interpretation also explains the asymmetry observed in the cross-model census (§5.5, Table 4). Habits such as explicit negation and anatomic checklist are general radiology conventions that any CXR-trained VLM is likely to encode, and indeed the boost directions transfer across the three architectures we tested (Jaccard 0.65–0.75, cosine >0.91). Template traps are different: each model has its own, because the templates a model defaults into are shaped by its own training rather than by general radiology conventions. RadVLM's particular end-of-report negation chain and its repetition loop on prosthetic-valve descriptions are not shared by LLaVA-Rad or CheXOne, and the suppress directions therefore do not transfer (Jaccard 0.41–0.58). The steering *recipe* is the same in every model; only the suppress *list* must be re-derived per model. Aggregated over the test set (Table 2), the steered model recovers more findings the reference contains (-654 missed-finding errors) but also adds findings the reference does not ($+1,112$ false-finding errors), exactly what we expect from a recipe that pushes the model toward more thorough reporting (boost) while removing template fragments that were sometimes real and sometimes fabricated (suppress).

Quantitative analysis of activation features. The qualitative patterns above, as boost features on radiology habits, suppress features on template traps, bring another question: *where in the generated report does each feature actually activate?*. We answer this by using the generated reports from unsteered RadVLM on 10,000 validation studies, and activated features (Activation > 2.0) on every token at layer 16. For each reported feature, Table 12 shows how many studies it activates on, how strongly it activates on average, and at which point in the report it activates most (early, middle, or late). By analyzing this, we observed three patterns as follows:

- **Features related to the repetition-loop are rare and late.** Feature #15810 as FF-boost activates on only 43 out of 9,951 studies, but when it activates, 87% of it occurs in the second half of the report. And all three of its top activation snippets in Table 11 are literal phrase replications. It is not a clinically related observation; it is a specific alarm for a particular failure mode that happens in generation during late steps.
- **"Good Reporting" features activate everywhere.** Features like #5872 and #1925 (both FF-boost) fire on every single validation study. #5872 has the highest average activation ($\bar{a}{=}8.43$). These are the broad "write a proper radiology report" features that our boost-side amplification supports.
- **Template-trap features concentrate at the end.** Feature #4240 (WL-suppress) fires 98% of the time in the second half of the report, with a median position of 0.92 (nearly the last token). This matches the canonical end-of-report negation list ("*No pleural effusions. No pulmonary edema. No pneumonia.*") as a fixed phrase the model appends by default, regardless of image content. Zeroing it produces more wrong-location errors ($\Delta_{\mathrm{WL}} = -0.09$), which is exactly what we would expect if the feature controls an often-useful but sometimes-incorrect template.

Figure 4 shows the full decode-position distributions for all 12 reported features. In short, each feature activates at the part of the report where its effect matters: boost features that prevent repetition are most active toward the end, where repetition occurs; suppress features tied to standard phrasing are active wherever that phrasing appears. The position patterns are consistent with the causal roles, providing evidence that the SAE has learned an interpretable structure rather than arbitrary directions.

Feature	Role / target (Δ)	Hypothesised pattern	Act.	Generated-report context (… window around the most-active position)
#31611	**supp** / $\Delta_{\mathrm{FF}} = -0.16$	*support-device language*	10.39	…ead tip in the distal SVC. Heart size is upper limits of normal. There is mild prominence of the pul…
			10.21	…e with distal lead tip in the proximal right atrium. There is a right-sided Swan-Ganz catheter with…
			10.18	…e with the distal lead tip in the distal SVC. There is an enteric tube whose side port is at the GE…
#15810	**boost** / $\Delta_{\mathrm{FF}} = +0.30$	*repetition-loop brake*	18.42	…is a prosthetic aortic valve. There is a prosthetic aortic valve. There
			18.31	…is a prosthetic aortic valve. There is a prosthetic aortic valve. There is a prosthetic aortic valv…
			18.28	…is a prosthetic aortic valve. There is a prosthetic aortic valve. There is a prosthetic aortic valv…
#2541	**supp** / $\Delta_{\mathrm{MF}} = -0.09$	*paired-finding listing*	16.86	…e left pleural effusion with associated left basilar atelectasis. There is no pneumothorax. The righ…
			16.81	…ht greater than left. There is pulmonary vascular congestion and mild pulmonary edema. The cardiac s…
			16.80	…e left pleural effusion with associated left basilar atelectasis. The right lung is clear. There is…
#1925	**boost** / $\Delta_{\mathrm{MF}} = +0.37$	*incidental findings*	10.02	…reflect areas of atelectasis. No pneumothorax is identified. Multiple old bilateral rib fractures a…
			9.96	…A vagal nerve stimulator is present. Multiple old right rib fractures are noted.
			9.81	…al. A vagal nerve stimulator is present. Multiple old right rib fractures are noted.
#4240	**supp** / $\Delta_{\mathrm{WL}} = -0.09$	*negation-chain template*	9.52	…tion. Moderate cardiomegaly. No pleural effusions. No pulmonary edema. No pneumonia.
			9.40	…rdiomegaly. No pleural effusions. No pulmonary edema. No pneumonia.
			9.36	…ax. No pleural effusions. No pulmonary edema.
#28456	**boost** / $\Delta_{\mathrm{WL}} = +0.06$	*anatomic detail*	7.64	…umothorax. There are low lung volumes. Bibasilar opacities are a combination of atelectasis and smal…
			6.96	…x. There are low lung volumes. Bibasilar opacities are a combination of atelectasis and small effusi…
			6.84	…C. There is no pneumothorax. Bibasilar opacities are a combination of small effusions and adjacent a…
#29962	**supp** / $\Delta_{\mathrm{WS}} = -0.11$	*generic-normal template*	7.94	…is normal. The aorta is tortuous. There is no pneumothorax or pleural effusion. There is no focal c…
			7.87	…ildly enlarged. The aorta is tortuous. There is no pleural effusion or pneumothorax. The lungs appea…
			7.83	…ildly enlarged. The aorta is tortuous. There is no pleural effusion or pneumothorax. The lungs appea…
#7046	**boost** / $\Delta_{\mathrm{WS}} = +0.05$	*precise laterality*	13.62	…ve subcutaneous emphysema in the right neck and chest wall. There is no pneumothorax. There are mode…
			13.61	…f subcutaneous emphysema in the right chest wall. There is a small amount of atelectasis at the righ…
			13.58	…omediastinal silhouette is within normal limits. Right chest wall port is seen with catheter tip in…
#5872	**boost** / $\Delta_{\mathrm{FF}} = +0.28$	*anatomic checklist*	20.83	The lungs are clear. The cardiomediastinal silhouette is within normal…
			20.72	The lungs are clear. The cardiomediastinal silhouette is within normal…
			20.71	The lungs are clear. The cardiomediastinal silhouette is within normal…
#16965	**boost** / $\Delta_{\mathrm{FF}} = +0.27$	*explicit negation*	7.83	…idence of pneumothorax. The lung volumes are normal. Normal size of the cardiac silhouette. Normal h…
			7.56	There is no evidence of pneumothorax. There is a small left pleural effusion. There is no…
			7.54	There is no evidence of pneumothorax. There is no evidence of pulmonary edema. There is no…
#20872	**boost** / $\Delta_{\mathrm{MF}} = +0.23$	*anatomic checklist (osseous)*	10.43	The lungs are clear. There is no pneumothorax or pleural effusion. The osseous…
			8.92	The lungs are clear. The cardiomediastinal silhouette…
			8.55	The lungs are clear. The cardiomediastinal silhouette is within normal…
#8368	**supp** / $\Delta_{\mathrm{WS}} = -0.10$	*Late-report repetition / devices*	10.28	…s a prosthetic pulmonary valve. There is a prosthetic pulmonic valve. There is a prosthetic pulmonic…
			9.92	…There is a nasogastric tube whose distal tip and side port are below the GE junction. There is a lef…
			9.67	…here is mild pulmonary edema. There is atelectasis at the lung bases.

Table 11: **Generated-report contexts where reported SAE features respond most strongly (RadVLM, layer 16, MIMIC-CXR validation set).** For each reported feature from Table 3 and the per-layer top features (Table 8).

Feature	Role/Δ	#Active	#Stud.	$\bar{a}$	p_{50}^{pos}	frac. late	rep.
#31611	**supp**/FF -0.16	34,058	2,853	3.89	0.61	65%	1/3
#15810	**boost**/FF $+0.30$	932	43	5.67	0.76	87%	3/3
#2541	**supp**/MF -0.09	20,972	6,798	7.12	0.63	66%	0/3
#1925	**boost**/MF $+0.37$	167,720	9,951	3.22	0.43	44%	0/3
#4240	**supp**/WL -0.09	2,203	825	4.08	0.92	98%	0/3
#28456	**boost**/WL $+0.06$	47,526	4,952	2.78	0.49	50%	0/3
#29962	**supp**/WS -0.11	2,321	390	2.85	0.65	76%	1/3
#7046	**boost**/WS $+0.05$	917	502	6.28	0.63	69%	0/3
#5872	**boost**/FF $+0.28$	43,626	9,951	8.43	0.66	65%	0/3
#16965	**boost**/FF $+0.27$	863	290	4.25	0.61	70%	1/3
#20872	**boost**/MF $+0.23$	440	152	3.05	0.17	25%	0/3
#8368	**supp**/WS -0.10	89,383	7,294	3.47	0.76	87%	1/3

Table 12: **Activation profile of reported features at layer-16 on the RadVLM per-token validation pool** (10,000 decodes; distinct `study_id` counts in #Stud.; pre-activation threshold > 2.00; causal Δ uses $N{=}100$ studies per feature, §3). #Active = number of token positions above threshold; #Stud. = distinct study ids on which the feature is active; $\bar{a}$ = mean pre-activation when active; p_{50}^{pos} = median normalised decode position (0 = first token of the report, 1 = last); frac. late = fraction of activations in the second half of the report; rep. = number of top-3 activation-context snippets (Table 11) that contain a literal phrase repetition.

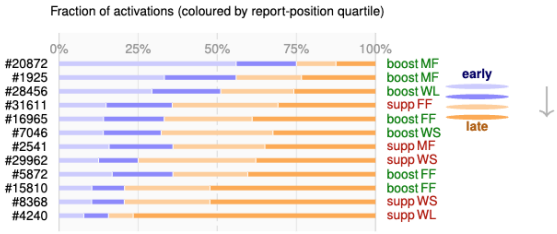


Figure 4: Where each reported layer-16 feature fires in the generated report (10,000 validation decodes). Each stacked bar shows what fraction of a feature's total activations fall in each position quartile: early [0, .25), early-mid [.25, .50), late-mid [.50, .75), late [.75, 1). Boost features (e.g. #15810) concentrate in the late quartiles; suppress features spread more evenly. Sorted by median position (bottom = latest).

E Cross-model feature analysis: formal definition

This section gives the exact recipe behind the cross-model census reported in Table 4. .

Setup. Let $\mathcal{T} = \{\text{FF}, \text{MF}, \text{WL}, \text{WS}\}$ denote the four GREEN error types we steer. For each model $m \in \{\text{RadVLM}, \text{LLaVA-Rad}, \text{CheXOne}\}$, layer $\ell \in \{8, 16, 20, 24\}$ and direction $d \in \{\text{SUPPRESS}, \text{BOOST}\}$, the per-error-type causal screen of §3 (run with `green_pertype` scoring) returns, for each candidate feature f, the 4-vector

$$\Delta_f^{(m,\ell)} = \left(\Delta_{\text{FF}}, \Delta_{\text{MF}}, \Delta_{\text{WL}}, \Delta_{\text{WS}}\right)_f \in \mathbb{R}^4 \quad (6)$$

where $\Delta_{t,f} = \mathbb{E}_x\left[\#\,\text{err}_t(y^{\text{abl}_f}(x)) - \#\,\text{err}_t(y^{\text{base}}(x))\right]$ is the mean change in t-type error count when feature f is zeroed at layer ℓ, averaged over the validation samples that passed the word-F1 prefilter. The causal-screen output also ships pre-ranked, sign-consistent target lists $T_t^{(m,\ell,d)} = \texttt{intervention_targets}[d][t]$, ordered by $|\Delta_{t,f}|$ with sign matched to d (suppress: $\Delta_t < 0$; boost: $\Delta_t > 0$).

Consensus set. For each (m, ℓ, d) we form a deterministic consensus list $C^{(m,\ell,d)}$ of size $N = 100$ by round-robin-merging the four ranked per-type lists, taking the first occurrence of each feature:

$$C^{(m,\ell,d)} = \text{RR}(T_{\text{FF}}^{(m,\ell,d)}, T_{\text{MF}}^{(m,\ell,d)}, T_{\text{WL}}^{(m,\ell,d)}, T_{\text{WS}}^{(m,\ell,d)})_{1:N} \quad (7)$$

This gives every error type proportional representation while preserving the per-type rank order. Sizes n_a, n_b in Table 4 report $|C^{(a,\ell,d)}|, |C^{(b,\ell,d)}|$ (smaller than N when fewer than N features pass the screen, e.g. a thinly-populated boost direction on CheXOne at $\ell = 8$).

Two summaries per (model, layer, direction). Each model's behaviour at (ℓ, d) is summarised by two 4-vectors derived from $C^{(m,\ell,d)}$. The *signature vector* is the mean per-feature causal effect over the consensus set:

$$\mathbf{s}^{(m,\ell,d)} = \frac{1}{|C^{(m,\ell,d)}|} \sum_{f \in C^{(m,\ell,d)}} \Delta_f^{(m,\ell)} \in \mathbb{R}^4. \quad (8)$$

The *categorical profile* is the empirical distribution of dominant, direction-consistent error types across the consensus set:

$$\pi_t^{(m,\ell,d)} = \frac{1}{V} \sum_{f \in C^{(m,\ell,d)}} \mathbf{1}\left[t = \arg\max_{t' \in \mathcal{T} : \text{sgn}(\Delta_{t',f}) = \text{sgn}(d)} |\Delta_{t',f}|\right] \quad (9)$$

where $V = \sum_f \mathbf{1}[\exists\, t' : \text{sgn}(\Delta_{t',f}) = \text{sgn}(d)]$ is the number of features that have at least one direction-consistent component, and $\text{sgn}(\text{SUPPRESS}) = -1$, $\text{sgn}(\text{BOOST}) = +1$

Pairwise similarities. For an unordered model pair (a, b) and a given (ℓ, d), the two reported numbers in Table 4 are

$$\text{signature_cosine}_{a,b} = \frac{\mathbf{s}^{(a)} \cdot \mathbf{s}^{(b)}}{\|\mathbf{s}^{(a)}\|_2 \, \|\mathbf{s}^{(b)}\|_2} \in [-1, 1] \quad (10)$$

and the Ruzicka (weighted-Jaccard) similarity of the categorical profiles,

$$\text{functional_jaccard}_{a,b} = \frac{\sum_{t\in\mathcal{T}} \min\left(\pi_t^{(a)}, \pi_t^{(b)}\right)}{\sum_{t\in\mathcal{T}} \max\left(\pi_t^{(a)}, \pi_t^{(b)}\right)} \in [0,\ 1] \tag{11}$$

Cosine in Eq. 10 compares the *direction and magnitude profile* of the two models' causal effects in $\mathcal{T}$-space; Ruzicka in Eq. 11 compares the discrete *which-error-type-dominates* distribution. Both are basis-free and therefore well defined across architectures that do not share SAE indices.

Aggregation and CIs. Table 4 reports, per direction and per model pair, the mean of (10) and (11) over the four steered layers $\ell \in \{8, 16, 20, 24\}$. The bracketed 95% confidence intervals are percentile bootstrap CIs over those four layers ($B = 10^4$ resamples). Boost-vs-suppress contrasts use a separate paired bootstrap on the difference of layer means.

F Cross-dataset transfer: IU-Xray

The SAEs and identifies features in the paper are based on MIMIC-CXR. To test whether the steering recipe transfers across datasets, not just across model architectures, we evaluate the *same* SAE weights, feature lists, and operating point ($\alpha = 0.20$, $K = 20$, layers $\{8, 16, 20, 24\}$, Combined mode) on **IU-Xray**, a publicly available collection of $3,955$ frontal chest radiographs with paired radiology reports from Indiana University. After filtering for image availability, $N = 3,307$ studies are decoded with unsteered and SAE-steered RadVLM.

Table 13 shows the results of the steered-RadVLM on IU-Xray. We also report the per-error-type breakdown on IU-Xray. The dominant pattern matches MIMIC-CXR (Table 2): missing findings drop sharply (-850) while false findings increase moderately ($+223$). Unlike MIMIC-CXR, total significant errors *decrease* on IU-Xray (-713), because the larger MF reduction on shorter reports more than offsets the FF increase.

G Quantitative examples: Improvement, regression

Table 15 shows four IU-Xray cases annotated with GREEN error tags derived directly from GREEN model raw completions. IU-Xray is a public dataset, so we reproduce the original chest radiographs and full report text. The steering removes a false cardiomegaly report on an otherwise normal study (iuxray_002410, GR: 0.67→1.00); recovers a hemodialysis catheter and its tip position that the baseline omitted entirely (iuxray_000286, GR: 0.17→0.80); and nudges a borderline-enlarged heart from "within normal limits" to "top normal in size", though the RLL granuloma remains unreported (iuxray_001563, GR: 0.00→0.75). The regression case (iuxray_001044) shows the FF↑ trade-off: the baseline accurately describes a subclavian catheter and a normal study (GR=1.00); steering hallucinates a pleural effusion and atelectasis (GR: 1.00→0.20).

Setting	GREEN↑	RG↑	CXB$_\mu$↑	BS↑	BL↑	RL↑	Comp.↑
RadVLM (Qwen3-VL 8B) on IU-Xray — α=0.20, K=20							
Base	49.7	24.7	49.2	88.6	5.6	26.6	46.5
+ SAE Combined	**53.6**	**28.9**	47.0	**88.7**	**6.5**	**26.7**	**49.0**
Δ *(pp)*	*+3.8*	*+4.1*	*−2.2*	*+0.2*	*+0.9*	*+0.1*	*+2.5*
rel. %	*+7.7%*	*+16.7%*	*−4.5%*	*+0.2%*	*+15.7%*	*+0.3%*	*+5.4%*

Table 13: RadVLM headline results on **IU-Xray** ($N = 3,307$ reports), using the same SAE weights, feature lists, and operating point ($\alpha = 0.20$, $K = 20$, layers $\{8, 16, 20, 24\}$) as the MIMIC-CXR test in Table 1. GREEN, RadGraph, BERTScore, BLEU-4, and ROUGE-L all improve; CheXbert 14-label micro F1 drops slightly (-4.5% rel.), but the net Composite gain is $+2.5$ pp ($+5.4\%$ rel.). Column definitions match Table 1.

GREEN error type	Baseline	+ SAE Combined	Δ
False finding (FF)	2,140	2,363	+223
Missing finding (MF)	6,498	5,648	**−850**
Wrong location (WL)	203	279	+76
Wrong severity (WS)	310	245	−65
False comparison (FC)	299	221	−78
Missing comparison (MC)	115	96	−19
Total significant errors	**9,565**	**8,852**	**−713**

Table 14: Per-error-type breakdown on **IU-Xray**, at the same operating point as Table 13. The dominant pattern matches MIMIC-CXR (Table 2): missing findings drop sharply (-850) while false findings increase moderately ($+223$). Unlike MIMIC-CXR, total significant errors *decrease* on IU-Xray (-713), because the larger MF reduction on shorter reports more than offsets the FF increase.

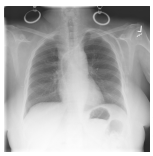
Case	CXR	Ground truth	Baseline (unsteered)	+SAE steering	GREEN (bl.→st.)
Improvement	iuxray_002410 normal study	The cardiomediastinal silhouette is within normal limits for size and contour. The lungs are normally inflated without evidence of focal airspace disease, pleural effusion, or pneumothorax. Osseous structures are within normal limits for patient age.	The lungs are clear without focal consolidation **[OK]**. No pleural effusion or pneumothorax is seen **[OK]**. The cardiac silhouette is mildly enlarged **[FF]**. No pulmonary edema is seen **[INSIG]**.	The lungs are clear without focal consolidation, pleural effusion, or pneumothorax **[OK]**. The cardiomediastinal silhouette is within normal limits **[OK]**. The visualized osseous structures are unremarkable **[OK]**.	0.67 → 1.00
Improvement	 iuxray_000286 hemodialysis catheter	A left-sided hemodialysis catheter is in place with its distal tip at the right atrium. The cardiac silhouette and mediastinal contours are within normal limits. There is no focal opacity. There is no pneumothorax. No large pleural effusion.	The lungs are clear **[FF]**. The cardiomediastinal silhouette is within normal limits **[OK]**. No acute osseous abnormalities.	There is a left-sided central venous catheter with the tip in the right atrium **[MF]**. The lungs are clear without focal consolidation, pleural effusion, or pneumothorax **[OK]**. The cardiomediastinal silhouette is normal **[OK]**.	0.17 → 0.80
Improvement	iuxray_001563 borderline heart, RLL granuloma	Heart size borderline enlarged. No focal alveolar consolidation, no definite pleural effusion seen. No typical findings of pulmonary edema. Dense nodule in the right lower lobe suggests a previous granulomatous process.	The lungs are clear. There is no pneumothorax. The heart and mediastinum are within normal limits **[FF]**. Regional bones and soft tissues are unremarkable.	The lungs are clear without focal consolidation, pleural effusion or pneumothorax **[OK]**. The pulmonary vasculature is not engorged. The cardiac silhouette is top normal in size **[WS]**. The mediastinal and hilar contours are within normal limits. No acute osseous abnormality is detected.	0.00 → 0.75
Regression	iuxray_000779 slightly large heart	The heart is slightly large. Pulmonary vasculature is normal. No infiltrates.	The lungs are clear without focal consolidation **[OK]**. No pleural effusion or pneumothorax is seen. The cardiac silhouette is mildly enlarged **[OK]**. The aorta is calcified and tortuous.	The lungs are clear without focal consolidation, pleural effusion, or pneumothorax. The cardiac silhouette is enlarged **[OK]**. The mediastinal and hilar contours are within normal limits. There is a mild compression deformity of a lower thoracic vertebral body **[FF]**.	1.00 → 0.50

Table 15: **IU-Xray qualitative steering cases** (RadVLM, Combined boost+suppress, α=0.20, K=20, layers $\{8, 16, 20, 24\}$. IU-Xray is a public dataset, so images and reports can be reproduced here. Cases span false cardiomegaly on a normal study, device recovery, borderline cardiomegaly with granuloma, and a moderate regression on an enlarged-heart study. Red + bracket tags mark significant errors: **FF** false finding, **MF** missing finding. Teal + **OK** marks findings that match the reference. Orange marks residual significant issues in the steered output. Gray **INSIG** / **WS** mark clinically insignificant wording differences (not counted toward GREEN). All tags are derived from GREEN model raw completions. **Improvement** (iuxray_002410): steering removes a false cardiomegaly report, reaching GREEN = 1.00. **Improvement** (iuxray_000286): the baseline omits the hemodialysis catheter; steering recovers the device and tip position (GREEN = 0.80). **Improvement** (iuxray_001563): baseline falsely calls the heart "within normal limits" on a borderline-enlarged study with an RLL granuloma; steering moves heart size to "top normal" but still omits the nodule (GREEN = 0.75). **Regression** (iuxray_000779): on a study with a slightly large heart, steering retains cardiomegaly but hallucinates a vertebral compression deformity (GREEN = 1.00→0.50), illustrating the FF trade-off (Table 14).