IF MEDICAL CONCEPTS ARE STILL THERE, WHY DO MED-LVLMS GET THEM WRONG?

Anonymous authors
Paper under double-blind review

ABSTRACT

Why do medical large vision–language models (Med-LVLMs) get medical concepts wrong when those concepts remain recoverable from their internal representations? We investigate this gap between concept availability and answer correctness by tracing medical concepts through the visual encoder, vision–language connector, and language decoder. Linear probes reveal no sustained loss of concept decodability, while representational analyses indicate that encoder features are preserved across the connector. The decoder, however, exhibits differences in how available information is used: incorrect answers show weaker attention to target regions, and correct and incorrect answers become increasingly separable in middle-layer MLP activations. Activation patching further identifies middle-to-late MLP components whose clean outputs restore support for concept presence after target-region masking. These findings suggest that errors can arise from a failure to use available medical information, rather than from its loss alone. Guided by this evidence, we introduce MEDSTEER, which uses concept-specific directions derived from original and target-masked images to edit decoder activations, with a presence probe gating each edit. Across three Med-LVLMs, MEDSTEER improves average VQA accuracy by 5.2–12.6 percentage points over the unedited models. Our results identify a gap between representing medical concepts and using them to answer correctly, and demonstrate that concept-guided decoder editing can help bridge this gap.

1 INTRODUCTION

Medical large vision-language models (Med-LVLMs) support applications such as medical visual question answering and image interpretation (Li et al., 2023; Xu et al., 2026; Chen et al., 2024; Chang et al., 2026). Despite their progress, these models can generate inaccurate or hallucinated answers that are inconsistent with the supplied image (Xia et al., 2024; Gu et al., 2024; Chang et al., 2025). Such failures may arise at different stages: the visual encoder may not recognize a relevant medical concept, the vision–language connector may discard it, or the language decoder may receive the evidence but fail to use it. Existing mitigation methods intervene through contrastive decoding (Leng et al., 2024), attention-aware generation (Huang et al., 2024), or representation steering (Liu et al., 2025; Sun et al., 2025). Although effective, these methods generally do not establish where concept information is preserved, where its use breaks down, or which components provide principled intervention sites.

We address this gap by tracing medical concepts, including anatomical structures, abnormalities, and imaging findings, through the complete Med-LVLM pipeline. Using concepts annotated for both presence and spatial extent, we ask three questions: *Is concept information preserved as it enters the language model? Through which decoder components does it influence the answer? Can these components be steered to improve concept recognition?*

Tracing concept information through Med-LVLMs. We study these questions in three representative Med-LVLMs: MedSIGHT (Chang et al., 2026), HuatuoGPT-Vision (Chen et al., 2024), and Lingshu (Xu et al., 2026). We first train linear probes on representations from the visual encoder, vision-language connector, and every decoder layer (Section 4). Concept decodability remains consistently high, with no sustained transition from strong to poor performance at any stage. We then directly compare encoder and connector representations using subspace similarity and linear recon-

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

struction (Section 5). The connector preserves both concept recoverability and much of the dominant encoder feature structure. These results provide little evidence that the connector is a major information bottleneck for the evaluated concepts. Instead, they reveal a more subtle failure mode: the model may retain concept information internally while still producing an incorrect answer.

From information availability to causal influence. The presence of decodable information does not imply that the decoder uses it. We therefore examine target-region attention and MLP activations in the language decoder (Section 6). Incorrect answers tend to exhibit weaker attention to the annotated concept region, while correct and incorrect examples become increasingly separable in MLP activation space across the middle decoder layers. Because these observations are correlational, we complement them with activation patching. We mask the target region and restore individual attention or MLP outputs from the corresponding unmasked run. The strongest recovery of concept-presence support is concentrated in a small number of middle-to-late MLP layers. This intervention identifies specific components through which visual evidence can causally influence the answer.

From causal localization to activation steering. Guided by these findings, we introduce MEDSTEER, a simple yet effective probe-guided activation-steering method (Section 7). For each concept, MEDSTEER constructs a presence direction by averaging the activation differences between original and target-masked positive training examples at a component selected through activation patching. Because adding this direction unconditionally could bias the model toward positive answers, we use an internal presence probe to gate the intervention. The direction is added when the probe predicts that the concept is present, subtracted when it predicts absence.

Our contributions are: (1) a framework for tracing medical concepts across the visual encoder, connector, and language decoder; (2) evidence that target concepts remain recoverable and structurally preserved even when the model answers incorrectly; (3) causal localization of middle-to-late decoder components through which visual evidence affects concept-presence predictions; and (4) MEDSTEER, a probe-guided intervention that improves or preserves concept-recognition accuracy without updating model parameters.

2 RELATED WORK

Medical LVLMs support image interpretation and visual question answering (Li et al., 2023; Moor et al., 2023; Chen et al., 2024; Chang et al., 2026), but remain susceptible to hallucination and other reliability failures (Xia et al., 2024; Gu et al., 2024; Chang et al., 2025). Existing mitigation methods use contrastive decoding (Leng et al., 2024; Mahdavi et al., 2026; Liu et al., 2024), or attention-aware generation (Huang et al., 2024). Complementary work uses linear probes to measure information recoverable from frozen representations (Alain & Bengio, 2017; Belinkov, 2022), although decodability does not establish that the information affects model predictions (Hewitt & Liang, 2019; Ravichander et al., 2021). Recent studies further examine visual information across VLM encoders, connectors, and decoders (Chandhok et al., 2025; Neo et al., 2024; Jiang et al., 2025). Activation steering offers a means of modifying model behavior through internal representations (Turner et al., 2023; Rimsky et al., 2024). Recent VLM approaches steer visual or textual states, condition interventions on visual neglect, or combine steering with logit adjustment (Liu et al., 2025; Sun et al., 2025; Li et al., 2025). Building on these lines of work, we combine probing and representational analysis with activation patching (Zhang & Nanda, 2024) to distinguish information availability from causal use. We then derive concept-specific directions from original and target-masked images and apply them through steering.

3 PRELIMINARIES

Many medical questions concern a specific anatomical structure, abnormality, imaging finding, or clinical object, which we collectively call a *medical concept*. Correctly answering such questions requires the concept to be recognized by the visual encoder, preserved by the vision–language connector, and used by the language decoder. We study this process through *medical concept tracing*, which follows information about a target concept from the visual encoder to the generated answer. We analyze three representative Med-LVLMs: MedSIGHT (Chang et al., 2026), HuatuoGPT-Vision-7B (Chen et al., 2024), and Lingshu-7B (Xu et al., 2026). Together, they cover two common vision–language interfaces: CLIP-based MLP projectors and Qwen-style patch mergers. All mod-

els remain frozen throughout our analysis. We construct concept-presence VQA examples from datasets included in BiomedParse (Zhao et al., 2024), whose segmentation labels provide both concept presence and location. The evaluation spans CT datasets for abdominal organs, renal tumors, and lung nodules (Ji et al., 2022; KiTS23 Challenge Organizers, 2023; Armato et al., 2011); MRI datasets for abdominal organs, cardiac structures, and brain tumors; CXR datasets for pneumothorax, lungs, COVID-19 infection, and lung opacity (Filice et al., 2020); and pathology, fundus, OCT, dermoscopy, and endoscopy images. Ground-truth answers are derived from the segmentation labels, with equal numbers of positive and negative examples sampled for each concept. Appendix A provides full dataset and construction details.

4 Do Medical Concepts Degrade Across Layers?

We first examine whether concept information is preserved as it propagates through a Med-LVLM. Specifically, we ask: *Does a medically relevant concept cease to be decodable at a particular stage of the model?* Intuitively, a Med-LVLM should preserve information about the concepts required to answer a given question. Suppose a concept is reliably decodable at layer $\ell - 1$, becomes poorly decodable at layer ℓ, and remains poorly decodable in subsequent layers. We refer to layer ℓ as a potential *concept degradation point*, because this transition would suggest that the layer transforms, suppresses, or discards concept information needed to produce the correct answer.

To identify potential concept degradation points, we train **linear probes** to predict the presence of a target concept from the representation at each stage of the model. Linear probes test whether a target property can be recovered by a simple classifier from frozen representations (Alain & Bengio, 2017; Belinkov, 2022). Prior work has used them to examine factual information encoded in LLM hidden states (Marks & Tegmark, 2024) and visual information represented across the encoder, connector, and decoder of VLMs (Chandhok et al., 2025). We apply the same family of linear readouts throughout the Med-LVLM pipeline. We index the visual encoder representation by $\ell = -1$, the connector representation by $\ell = 0$, and the outputs of the L LLM blocks by $\ell = 1, \ldots, L$. This unified indexing allows us to trace concept decodability from visual perception through vision–language alignment and language-model reasoning.

Let $\mathbf{h}_\ell(x, q(c)) \in \mathbb{R}^{d_\ell}$ denote the representation at layer ℓ, where x is the input image and $q(c)$ is a question about the presence of medical concept c. At the visual encoder ($\ell = -1$) and connector ($\ell = 0$), we mean-pool the corresponding spatial visual tokens following (Chandhok et al., 2025). These representations depend only on the image and are therefore independent of the textual query. At each LLM layer $\ell \in \{1, \ldots, L\}$, we extract the residual-stream representation at the final prompt position, after layer ℓ. The query $q(c)$ is held fixed across all examples for a given concept. For each concept c and layer ℓ, we define the probe as $\widehat{p}_c^{(\ell)}(x) = \sigma\Big(\big(\mathbf{w}_c^{(\ell)}\big)^\top \mathbf{h}_\ell(x, q(c)) + b_c^{(\ell)}\Big)$, where σ denotes the sigmoid function. We independently fit the probe parameters $\mathbf{w}_c^{(\ell)} \in \mathbb{R}^{d_\ell}$ and $b_c^{(\ell)} \in \mathbb{R}$ for each model, dataset, and concept, while keeping the visual encoder, connector, and LLM frozen. We report balanced accuracy to account for class imbalance in individual train–test splits. Importantly, probe performance measures whether concept information is *recoverable* from a representation by a linear readout. It does not establish that the model itself uses this information when generating an answer (Ravichander et al., 2021).

Figure 1 shows concept decodability across the complete Med-LVLM pipeline, from the visual encoder and connector to every decoder layer. Across the three models, balanced accuracy remains consistently high: 0.924–0.949 at the visual encoder, 0.930–0.951 at the connector, and 0.880–0.940 across the decoder layers. Although the layer-wise trajectories differ across models, none exhibits a sustained transition from high to low decodability. We therefore find no clear concept degradation point under linear probing.

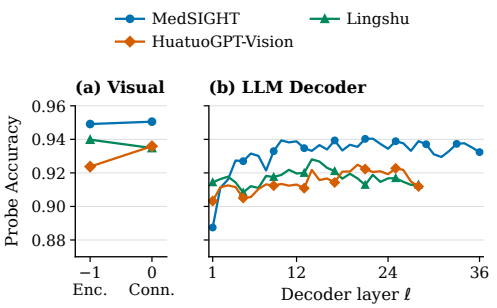


Figure 1: Concept decodability from encoder to decoder measured by test-set linear-probe accuracy, macro-averaged over modalities.

Although concept information remains linearly accessible throughout the network, this result

does not establish that it is faithfully preserved across component boundaries or used to generate the answer. We therefore next examine these two possibilities separately. We first test whether the vision–language connector preserves concept-relevant information as visual features enter the LLM; subsequent analyses investigate how the decoder uses this information during answer generation.

5 Beyond Decodability: Tracing Medical Concept Preservation

Layer-wise probing establishes that medical concepts remain broadly recoverable throughout the evaluated Med-LVLMs. However, similar probe performance at two stages does not necessarily imply that the underlying representations are preserved: a connector may substantially transform the feature space while retaining enough information for a linear classifier. We therefore examine the earliest multimodal interface, i.e., the vision–language connector, using complementary measures of concept recoverability and representational geometry.

A Med-LVLM first extracts visual features with an image encoder and then maps them into the LLM input space through a connector, such as an MLP projector or a patch-merging module. This transformation may compress, distort, or discard information relevant to the target medical concept. We therefore compare representations immediately before and after the connector. A substantial reduction in concept recoverability would implicate the connector as a potential information bottleneck, whereas stable recoverability would indicate that concept-relevant information remains accessible as the visual tokens enter the LLM.

We use SVCCA (Raghu et al., 2017) to measure the similarity between their dominant feature subspaces. We also report reconstruction R^2 (Kornblith et al., 2019), which measures how accurately a learned linear map can reconstruct the encoder representation from the connector output. These metrics assess whether the connector preserves the structure of the encoder representation. Formal definitions and implementation details are provided in Appendix B.

From the results shown in Table 1, we can observe that the representational similarity values are high: SVCCA ranges from 0.899 to 0.944, while encoder-reconstruction R^2 ranges from 0.870 to 0.962. These high values indicate that, despite mapping visual features into the LLM input space, the connector largely preserves the dominant structure of the encoder representations.

Table 1: Concept preservation results.

Model	SVCCA	$R^2(0\rightarrow-1)$
MedSIGHT	0.944	0.870
Lingshu	0.908	0.955
HuatuoGPT-Vision	0.899	0.962

These results, shown in Figure 1 and Table 1, provide little evidence that the connector is a major bottleneck for the medical concepts considered in our evaluation. The relevant information remains recoverable, and much of the encoder feature geometry can be reconstructed after projection. Consequently, failures cannot be readily attributed to the disappearance of target-concept information before it reaches the LLM. This finding shifts the analysis downstream: the central question is no longer whether the visual evidence reaches the decoder, but **whether and how the decoder uses that evidence to produce its answer**.

6 Tracing Visual Evidence Through the Language Decoder

6.1 Does the Decoder Attend to the Target Region?

To answer a concept-dependent question, the decoder must extract relevant evidence from the visual tokens. We therefore begin by examining whether its attention is concentrated on the annotated target region. We hypothesize that incorrect answers are associated with weaker attention to this region, suggesting the decoder may overlook relevant evidence even when it is present in the input.

We conduct this preliminary analysis on the CT validation set using MedSIGHT and Lingshu. Because the hidden state at the final prompt position is directly used to initiate answer generation, we measure its attention to the visual tokens corresponding to the ground-truth concept region. We average this attention across heads within each

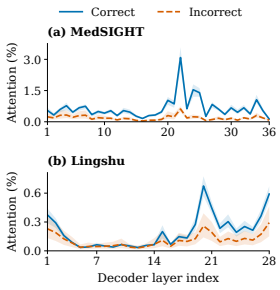


Figure 2: Attention to the target concept region.

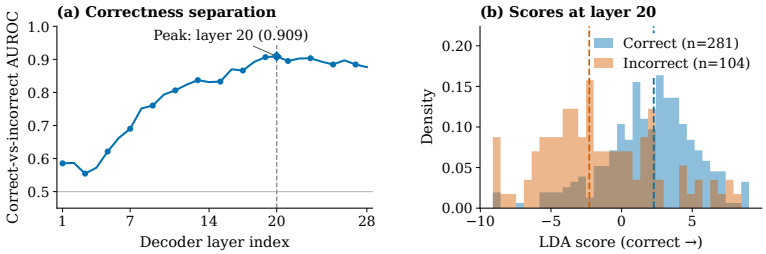


Figure 3: MLP activations differ between correct and incorrect answers in Lingshu. (a) AUROC of linear classifiers trained to distinguish correct from incorrect examples using the MLP output at each decoder layer. The dashed line marks the layer with the highest AUROC. (b) LDA projections of held-out layer-20 activations. Higher scores indicate greater alignment with the activation pattern associated with correct answers; dashed lines denote group medians.

decoder layer and compare the resulting layer-wise trajectories between correctly and incorrectly answered questions.

Figure 2 shows the head-averaged attention from the final prompt position to visual tokens corresponding to the annotated target region, stratified by answer correctness. We can observe that attention to the target region tends to be lower for incorrectly answered questions. Correct answers exhibit a pronounced attention peak in the middle-to-late decoder layers, whereas this peak is substantially weaker for incorrect answers. This pattern is consistent with prior findings that fine-grained multimodal integration often emerges in middle-to-late layers (Neo et al., 2024; Jiang et al., 2025). Full attention results of all three models are in Appendix C.1.

These results suggest that answer correctness is associated with how strongly the decoder attends to the relevant image region. However, attention weights alone do not establish that the attended information is incorporated into the representations used for prediction. We therefore next examine whether correct and incorrect answers are also distinguishable from the outputs of the decoder's MLP modules.

6.2 Do Decoder Activations Reflect Answer Correctness?

We analyze the MLP output at the final prompt position in each layer of Lingshu on the CT validation set. At each layer, we fit a linear classifier to distinguish examples that the model answers correctly from those it answers incorrectly. Classification performance measures the extent to which the two groups exhibit systematically different activation patterns; it does not, by itself, imply that these differences cause the model's predictions.

Figure 3(a) shows that correct and incorrect activations are only weakly separable in the early layers, with AUROC below 0.7. Separability increases through the middle layers and reaches approximately 0.9 around layer 20. Notably, this increase occurs near the attention peak observed in Figure 2. This correspondence is consistent with the possibility that differences in visual attention are propagated into subsequent MLP representations, although it does not establish a causal relationship between the two. To visualize this distinction, we fit linear discriminant analysis (LDA) on a probe-training subset and project held-out layer-20 activations onto the learned discriminant direction. As shown in Figure 3(b), correct answers generally receive higher projection scores than incorrect answers. Although the distributions overlap, their separation indicates that the MLP output contains information predictive of answer correctness. We include activation analysis on three models in Appendix C.2.

The attention and activation analyses identify decoder patterns associated with answer correctness, but both are correlational. To test whether particular component activations influence the answer, we next intervene on them directly.

6.3 Do Decoder Activations Causally Influence the Answer?

We hypothesize that some decoder components carry visual evidence that contributes to the model's prediction. We test this hypothesis using activation patching (Zhang & Nanda, 2024). For each example, we first mask the annotated target region to remove its visual evidence. We then restore

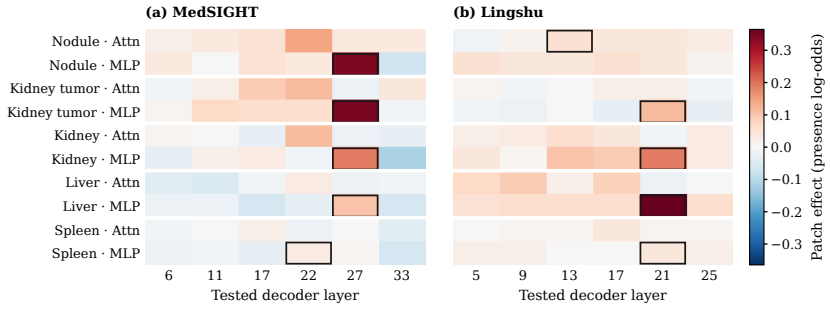


Figure 4: Effects of activation patching across decoder components. Each cell reports the mean recovery effect $\Delta S_{\ell,m}$ for concept c after restoring the attention or MLP output at the indicated layer. Boxed cells mark the component with the largest recovery effect for each concept within each model. Larger values indicate greater recovery of support for concept presence.

one decoder component output from the original, unmasked run and measure whether doing so recovers the model's support for concept presence.

For each sample, we first run the model on the original image x and a question $q(c)$ about concept c. We record the output $\mathbf{h}_{\ell,m}$ of each component $m \in \{\mathrm{Attn}, \mathrm{MLP}\}$ at decoder layer ℓ. We then mask the ground-truth region of c in x to obtain a corrupted image $\hat{x}$ and run the model on $(\hat{x}, q(c))$ to record the corresponding outputs $\hat{\mathbf{h}}_{\ell,m}$. Finally, for each component and layer in turn, we rerun the model on $(\hat{x}, q(c))$, replacing that component's output $\hat{\mathbf{h}}_{\ell,m}$ with its clean counterpart $\mathbf{h}_{\ell,m}$. All other components operate normally during the patched run. We measure how much each replacement restores the model's support for concept presence.

We quantify the model's support for concept presence using the log-odds $S(x, q(c)) = \log P(y^+ \mid x, q(c)) - \log P(y^- \mid x, q(c))$, where a larger value indicates stronger support for the presence of concept c, and $S(x, q(c)) > 0$ favors a positive answer. Let $\tilde{S}_{\ell,m}(\hat{x}, q(c))$ denote the corresponding score after restoring component (ℓ, m) in the target-masked run. We define the recovery effect as

$$\Delta S_{\ell,m}(x, q(c)) = \tilde{S}_{\ell,m}(\hat{x}, q(c)) - S(\hat{x}, q(c)). \tag{1}$$

A positive $\Delta S_{\ell,m}$ indicates that restoring component (ℓ, m) increases the model's support for concept presence relative to the unpatched target-masked run. For the concept-present examples used in this analysis, a larger positive value therefore indicates a larger increase in concept-presence support under this intervention.

We evaluate activation patching on five CT concepts and compare attention and MLP components at six selected decoder layers. As shown in Figure 4, the largest recovery effects are concentrated in a small number of middle-to-late MLP layers. For Lingshu, restoring the layer-21 MLP output produces a particularly strong recovery effect. This component lies immediately after the attention peak around layer 20 and near the layers where correct and incorrect MLP activations are most separable. This correspondence connects the preceding correlational observations to a causal intervention: restoring the output of a localized MLP component recovers support for concept presence even though the target region remains masked. The results therefore identify decoder components through which information removed by target-region masking can influence the answer. Additional results across three models are provided in Appendix C.3.

From Localization to Intervention

These analyses provide complementary evidence on how the decoder processes visual information. Attention to the target region and MLP activation patterns are associated with answer correctness, while activation patching identifies specific MLP components whose outputs causally affect concept-presence support. These findings motivate a stronger intervention question: *can we directly steer the identified components to improve the model's answers?*

In the next section, we test this possibility by editing decoder activations along a learned concept-presence direction.

Table 2: Main comparison across modalities on the test set. **Bold** indicates the best and underline the second best reported performance among comparable methods in each column, including ties.

Method	MedSIGHT					Lingshu					HuatuoGPT-Vision				
	CT	MRI	CXR	Others	Avg.	CT	MRI	CXR	Others	Avg.	CT	MRI	CXR	Others	Avg.
Unedited model	58.2	55.4	57.2	71.5	60.6	66.9	59.2	74.3	76.2	69.2	61.3	53.9	52.2	75.1	60.6
Activation Steering Baselines															
CAA	60.5	58.9	68.9	75.3	65.9	**70.7**	61.8	75.4	84.0	73.0	71.3	64.5	56.0	78.9	67.7
VTI	59.8	57.1	64.5	77.7	64.7	66.8	59.5	73.8	76.7	69.2	61.6	54.5	54.9	76.7	62.0
V-ITI	57.8	55.3	57.4	71.3	60.5	63.2	58.3	74.7	76.7	68.2	60.8	52.7	53.3	76.0	60.7
VISTA	58.4	54.5	59.0	71.5	60.8	66.9	60.7	74.3	76.2	69.6	61.7	53.9	52.2	75.1	60.7
ODESteer	58.3	55.2	57.3	71.7	60.6	67.9	60.1	75.0	77.3	70.1	66.3	53.7	52.2	75.2	61.9
Hallucination-Mitigation Baselines															
Med-VCD	59.5	53.3	62.8	77.2	63.2	66.3	59.6	74.2	76.8	69.2	61.9	55.8	52.7	75.9	61.6
M3ID	59.3	55.8	65.4	77.9	64.6	68.6	58.6	75.7	80.2	70.8	63.4	53.9	54.6	75.1	61.8
DoLa	57.7	56.3	61.8	71.5	61.8	70.1	61.0	75.0	75.8	70.5	70.1	**66.5**	53.0	81.9	67.9
PAI	59.3	55.8	63.0	78.9	64.3	64.4	57.4	75.7	81.6	69.8	63.7	53.2	58.2	75.7	62.7
MEDSTEER	**68.8**	**59.4**	**71.1**	**93.5**	**73.2**	69.4	**62.1**	**78.9**	**87.1**	**74.4**	**73.1**	64.3	**59.7**	**85.3**	**70.6**

7 CAN ACTIVATION STEERING IMPROVE CONCEPT RECOGNITION?

7.1 PROBE-GUIDED STEERING AT SELECTED COMPONENTS

For each model, activation patching identifies a component $(\ell^\star, m^\star)$ at which restoring the clean activation produces a strong recovery effect. Let $\mathbf{h}^i_{\ell^\star,m^\star}$ and $\hat{\mathbf{h}}^i_{\ell^\star,m^\star}$ denote the component outputs for the original and target-masked versions of positive training example i, respectively. Restoring the clean activation is equivalent to adding the difference $\mathbf{h}^i_{\ell^\star,m^\star} - \hat{\mathbf{h}}^i_{\ell^\star,m^\star}$ to the masked activation. We therefore estimate a concept-specific *presence direction* by averaging these differences over N_c positive training examples:

$$\mathbf{g}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} \left(\mathbf{h}^i_{\ell^\star,m^\star} - \hat{\mathbf{h}}^i_{\ell^\star,m^\star} \right). \tag{2}$$

This direction captures the activation change consistently induced by restoring visual evidence for concept c.

At test time, the target-region annotation is *unavailable*. Moreover, unconditionally adding $\mathbf{g}_c$ would systematically bias the model toward predicting that the concept is present. We therefore use an internal presence probe to gate the intervention. Let r_c denote the pre-sigmoid logit of the presence probe, computed from the unedited representation at layer $\ell^\star$. The threshold τ_c is in logit units; $\tau_c = 0$ corresponds to presence probability 0.5. We define the edited component output as follows:

$$\mathbf{h}'_{\ell^\star,m^\star} = \mathbf{h}_{\ell^\star,m^\star} + \alpha_c \eta_c \mathbf{g}_c, \qquad \eta_c = \begin{cases} +1, & r_c > \tau_c, \\ -1, & r_c < \tau_c, \\ 0, & r_c = \tau_c. \end{cases} \tag{3}$$

where α_c controls the intervention strength. The gate adds the presence direction when the probe indicates that the concept is present, subtracts it when the probe indicates absence. Although $\mathbf{g}_c$ is estimated from positive examples, the signed intervention assumes that its negative direction suppresses concept-presence evidence; we evaluate this design empirically through ablations below.

We select the concept-specific hyperparameters (α_c, τ_c) using the validation set, as detailed in Appendix D.2. Inference requires two forward passes: an initial unedited pass to compute r_c, followed by an edited pass in which Eq. (3) is applied. The intervention requires neither the ground-truth answer nor a target-region annotation at test time.

7.2 MAIN RESULTS ACROSS MODALITIES

We compare probe-guided steering with existing methods across CT, MRI, CXR, pathology, fundus, OCT, dermoscopy, and endoscopy. For each model and dataset, we learn a direction for each

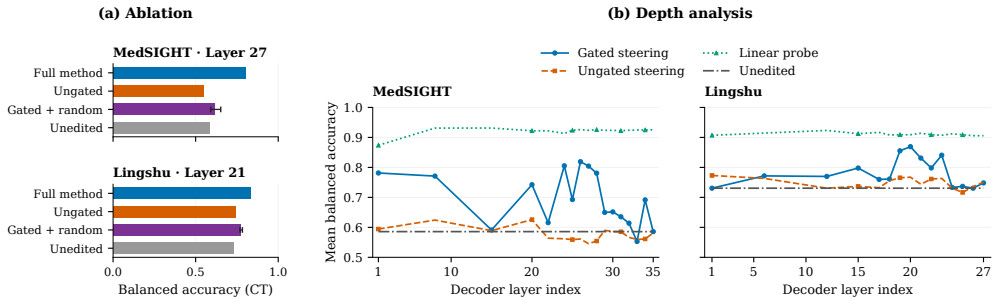


Figure 5: Analysis of probe-guided activation steering. (a) Ablations of the input-dependent gate and learned steering direction. (b) Performance across intervention depths for gated and ungated steering, with the original model and linear probe shown as references.

concept from 64 positive examples and evaluate it on the corresponding held-out test set. We report balanced VQA accuracy for each concept, macro-averaged across concepts and datasets within each modality group. Baselines include activation-steering methods: CAA (Rimsky et al., 2024), VTI (Liu et al., 2025), V-ITI (Sun et al., 2025), VISTA (Li et al., 2025), and ODESteer (Zhao et al., 2026); and hallucination-mitigation methods: Med-VCD (Mahdavi et al., 2026), DoLa (Chuang et al., 2024), M3ID (Favero et al., 2024), and PAI (Liu et al., 2024). For a controlled comparison, we adapt activation-steering baselines to concept-specific steering under the same protocol. Details are provided in Appendix D.5.

Table 2 shows that MEDSTEER improves balanced VQA accuracy across all model–modality pairs, with average gains of 12.6, 5.2, and 10.0 percentage points for MedSIGHT, Lingshu, and HuatuoGPT-Vision, respectively. CAA is the strongest baseline overall, yet MEDSTEER achieves higher average accuracy for all three models and leads in most modality comparisons. Other steering and hallucination-mitigation baselines, including contrastive decoding methods, show smaller gains overall. These results support the effectiveness of probe-guided intervention for concept recognition.

7.3 EFFECTIVENESS ANALYSIS

Are both the gate and direction necessary? We evaluate two ablations. First, *ungated steering* removes the input-dependent gate and applies the same signed edit to all examples. Second, *gated random steering* retains the gate but replaces $\mathbf{g}_c$ with a matched-norm random vector. Both variants reduce balanced accuracy relative to the complete method, as shown in Figure 5(a). These results support complementary roles for the two components: the gate determines the sign of intervention, while the learned direction determines how the activation is changed. Additional ablations across three models, including the number of training examples and steering strength α_c, are provided in Appendices D.3 and D.4.

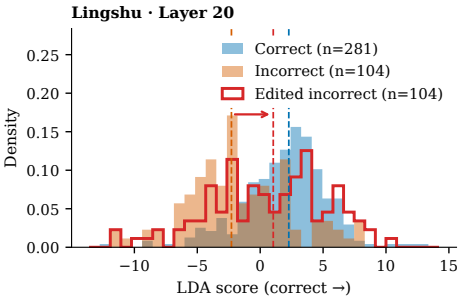


Figure 6: Activation shift induced by steering in Lingshu. Distribution of held-out activations projected onto the correctness-associated LDA direction before and after steering.

How sensitive is steering to intervention depth? We vary the intervention layer while holding the remaining settings fixed. Figure 5(b) compares gated steering, ungated steering, the original model, and linear-probe predictions across candidate layers. Gated steering produces little improvement in the early layers, whereas its largest gains occur in the middle-to-late layers. This depth dependence is consistent with the activation-patching results, which identify the same portion of the decoder as most influential for concept-presence support. Complete experimental settings and layer-wise results are provided in Appendix D.

How does steering change the activation state? Section 6.2 showed that correct and incorrect answers occupy partially separable regions of the activation space. We therefore ask whether steering moves missed-positive examples toward activation states associated with correct recognition.

For Lingshu, we here perform steering on-layer 20 MLP and project held-out layer-20 activations onto the correctness-associated LDA direction introduced in Section 6.2. As shown in Figure 6, steering shifts the median projection of incorrectly answered positive examples from -2.29 to 1.04, toward the median of 2.27 for correctly answered examples. This result indicates that the intervention moves missed-positive examples toward a representation associated with correct concept recognition. Because this analysis is correlational, it complements the causal evidence from the performance changes induced by steering.

7.4 Generalization Ability Analysis

Can concept directions transfer across datasets and modalities? Because the same medical concept can appear across datasets and imaging modalities, we ask whether its steering direction remains useful beyond the source setting. We evaluate CT-trained kidney and liver directions on SLAKE-CT, AMOS-MRI, and SLAKE-MRI, testing changes in dataset, modality, or both. Gates and strengths are adapted to each target. Both models improve across all three targets (Table 3), with the largest gains on AMOS-MRI. These results support transfer with target adaptation.

Table 3: Concept direction transfer.

Model	Target	n	Base	Steered
MedSIGHT	SLAKE-CT	155	73.1	**84.2**
	SLAKE-MRI	60	79.2	**83.9**
	AMOS-MRI	800	62.8	**85.0**
Lingshu	SLAKE-CT	155	90.2	**91.0**
	SLAKE-MRI	60	95.1	**96.9**
	AMOS-MRI	800	58.8	**90.1**

Can concept steering help open-ended generation? Recognizing a concept in a binary question does not guarantee that a model will describe it correctly in free-form text. We therefore evaluate concept steering on renal descriptions and IU-Xray reports. Specifically, we focus on transferring the concept direction for "lung opacity" from closed-ended VQA to open-ended generation. Target-term accuracy on renal descriptions improves by 6.0 and 1.5 percentage points for MedSIGHT and Lingshu, respectively (Table 4). On IU-Xray, target accuracy increases from 54.3% to 56.7% and from 51.0% to 51.9%, with small F1 gains across both all findings and non-target findings. These results suggest that steering can help models express target concepts beyond binary answers. The evaluation settings and examples are included in Appendix E.

Table 4: **Open-ended generation.** Base $\rightarrow$ steered scores (%). Acc: balanced accuracy; F1: micro F1.

Metric	MedSIGHT	Lingshu
Renal descriptions ($n = 200$)		
Target-term Acc	$41.0 \rightarrow$ **47.0**	$54.0 \rightarrow$ **55.5**
IU-Xray report ($n = 100$)		
Target Acc	$54.3 \rightarrow$ **56.7**	$51.0 \rightarrow$ **51.9**
F1 (14 findings)	$15.1 \rightarrow$ **15.9**	$23.5 \rightarrow$ **24.2**
F1 (non-target)	$12.3 \rightarrow$ **13.0**	$26.0 \rightarrow$ **26.2**

8 Conclusion

If medical concepts are still there, why do Med-LVLMs get them wrong? Our findings suggest that representing a concept does not guarantee that the model will use it to answer correctly. Concept information remains broadly recoverable through the visual encoder, connector, and decoder, while incorrect answers exhibit weaker attention to target regions and distinct middle-layer MLP activation patterns. Activation patching further shows that middle-to-late MLP components can restore concept-presence support after target-region masking. Guided by these findings, MedSteer uses probe-gated, concept-specific activation edits to improve average VQA accuracy by 5.2–12.6 percentage points across three Med-LVLMs. Transfer and open-ended evaluations suggest broader potential, although their exploratory nature limits generalization claims. Together, these results indicate that some errors reflect a gap between available medical information and its use in answering, and that targeted decoder editing can help narrow this gap.

AI USE STATEMENT

We used generative AI tools to assist with data preprocessing, code development, testing, and manuscript polishing. The research ideas were developed by the authors; generative AI was not used to originate the core research ideas. The authors retain responsibility for the methodology, analyses, results, and final manuscript, including all AI-assisted content and artifacts.

ETHICS STATEMENT

This work studies model reliability using existing medical imaging datasets. Its purpose is to understand and improve how Med-LVLMs use medical concept information, rather than to establish clinical readiness. Activation steering can introduce incorrect findings or suppress valid ones, and its effectiveness may vary across concepts, datasets, and patient populations. The reported benchmark gains do not establish clinical safety or equitable performance. Clinical use would require independent validation, assessment of potential harms and subgroup performance, and qualified medical oversight.

REPRODUCIBILITY STATEMENT

The main paper describes the concept-tracing analyses and probe-guided steering method. Appendix A describes dataset sources, concept definitions, and split construction; Appendices B and C detail the representation analyses and activation interventions. Steering protocols, hyperparameter selection, and ablations are described in Appendix D. Appendix E documents generation settings, scoring procedures, and post-hoc evaluation choices to clarify how the exploratory results were obtained.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR) Workshop*, 2017. arXiv:1610.01644.

Sharib Ali, Debesh Jha, Noha Ghatwary, et al. A multi-centre polyp detection and segmentation dataset for generalisability assessment. *Scientific Data*, 10:75, 2023. doi: 10.1038/s41597-023-01981-y. URL https://github.com/DebeshJha/PolypGen.

Samuel G. Armato, Geoffrey McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. *Medical Physics*, 38(2):915–931, 2011. doi: 10.1118/1.3528204. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3041807/.

Muhammad Naseer Bajwa, Gur Amrit Pal Singh, Wolfgang Neumeier, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection. *arXiv preprint arXiv:2006.09158*, 2020. URL https://arxiv.org/abs/2006.09158.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018. doi: 10.1109/TMI.2018.2837502. URL https://www.creatis.insa-lyon.fr/Challenge/acdc/.

Mateusz Buda, Ashirbani Saha, and Maciej A. Mazurowski. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in Biology and Medicine*, 109:218–225, June 2019. doi: 10.1016/j.compbiomed.2019.05.002. URL http://dx.doi.org/10.1016/j.compbiomed.2019.05.002.

Victor M. Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martin-Isla, Alireza Sojoudi, Peter M. Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, Mario Parreno, Alberto Albiol, Fanwei Kong, Shawn C. Shadden, Jorge Corral Acero, Vaanathi Sundaresan, Mina Saber, Mustafa Elattar, Hongwei Li, Bjoern Menze, Firas Khader, Christoph Haarburger, Cian M. Scannell, Mitko Veta, Adam Carscadden, Kumaradevan Punithakumar, Xiao Liu, Sotirios A. Tsaftaris, Xiaoqiong Huang, Xin Yang, Lei Li, Xiahai Zhuang, David Vilades, Martin L. Descalzo, Andrea Guala, Lucia La Mura, Matthias G. Friedrich, Ria Garg, Julie Lebel, Filipe Henriques, Mahir Karakas, Ersin Cavus, Steffen E. Petersen, Sergio Escalera, Santi Segui, Jose F. Rodriguez-Palomares, and Karim Lekadir. Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554, December 2021. doi: 10.1109/tmi.2021.3090082. URL http://dx.doi.org/10.1109/TMI.2021.3090082.

Shivam Chandhok, Wan-Cyuan Fan, Vered Shwartz, Vineeth N. Balasubramanian, and Leonid Sigal. Response wide shut? surprising observations in basic vision language model capabilities. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25530–25545, 2025. doi: 10.18653/v1/2025.acl-long.1241. URL https://aclanthology.org/2025.acl-long.1241/.

Aofei Chang, Le Huang, Parminder Bhatia, Taha Kass-Hout, Fenglong Ma, and Cao Xiao. MedHEval: Benchmarking hallucinations and mitigation strategies in medical large vision-language models. *arXiv preprint arXiv:2503.02157*, 2025. URL https://arxiv.org/abs/2503.02157.

Aofei Chang, Le Huang, Alex James Boyd, Parminder Bhatia, Taha Kass-Hout, Fenglong Ma, and Cao Xiao. MedSIGHT: Towards grounded visual comprehension in medical large vision-language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=a7mORMEWYX.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, et al. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Th6NyL07na.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1902.03368*, 2019. URL https://arxiv.org/abs/1902.03368.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. doi: 10.1093/jamia/ocv080. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5009925/.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14303–14312, 2024.

Ross W. Filice, Anouk Stein, Carol C. Wu, Veronica A. Arteaga, Stephen Borstelmann, Ramya Gaddikeri, Maya Galperin-Aizenberg, Ritu R. Gill, Myrna C. Godoy, Stephen B. Hobbs, et al. Crowdsourcing pneumothorax annotations using machine learning annotations on the NIH chest X-ray dataset. *Journal of Digital Imaging*, 33(2):490–496, 2020. doi: 10.1007/s10278-019-00299-9. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7165201/.

Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. Pannuke: An open pan-cancer histology dataset for nuclei instance segmentation and classification. In *Digital Pathology*, pp. 11–19. Springer International Publishing, 2019. doi: 10.1007/978-3-030-23937-4_2. URL http://dx.doi.org/10.1007/978-3-030-23937-4_2.

Zishan Gu, Changchang Yin, Fenglin Liu, and Ping Zhang. Medvh: Towards systematic evaluation of hallucination for large vision language models in the medical context. *arXiv preprint arXiv:2407.02730*, 2024.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pp. 2733–2743, 2019.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 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*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Huang_OPERA_Alleviating_Hallucination_in_Multi-Modal_Large_Language_Models_via_Over-Trust_CVPR_2024_paper.pdf.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, and Ping Luo. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25004–25014, 2025.

KiTS23 Challenge Organizers. The 2023 kidney and kidney tumor segmentation challenge: The data and data annotation process, 2023. URL https://kits-challenge.org/kits23/. Official source annotation definitions; accessed September 11, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 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*, pp. 13872–13882, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.pdf.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=7BKcLeHQsm.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1650–1654, 2021.

Sheng Liu, Haotian Ye, and James Y Zou. Reducing hallucinations in large vision-language models via latent space steering. In *International Conference on Learning Representations*, volume 2025, pp. 72402–72419, 2025.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pp. 125–140. Springer, 2024.

Zahra Mahdavi, Zahra Khodakaramimaghsoud, Hooman Khaloo, Sina Bakhshandeh Taleshani, Erfan Hashemi, Javad Mirzapour Kaleybar, and Omid Nejati Manzari. Med-vcd: Mitigating hallucination for medical large vision language models through visual contrastive decoding. *Computers in Biology and Medicine*, 200:111347, 2026.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pp. 353–367. PMLR, 2023.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. *arXiv preprint arXiv:2410.07149*, 2024. ICLR 2025.

José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, et al. REFUGE challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical Image Analysis*, 59:101570, 2020. doi: 10.1016/j.media.2019.101570. URL https://arxiv.org/abs/1910.03667.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, 2021.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of EMNLP*, pp. 1500–1519, 2020. doi: 10.18653/v1/2020.emnlp-main.117. URL https://aclanthology.org/2020.emnlp-main.117/.

Nan Sun, Zhenyu Zhang, Xixun Lin, Kun Wang, Yanmin Shang, Naibin Gu, Shuohuan Wang, Yu Sun, Hua Wu, Haifeng Wang, et al. V-iti: Mitigating hallucinations in multimodal large language models via visual inference-time intervention. *arXiv preprint arXiv:2512.03542*, 2025.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *Advances in Neural Information Processing Systems*, 37:140334–140365, 2024.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.

Hongjue Zhao, Haosen Sun, Jiangtao Kong, Xiaochang Li, Qineng Wang, Liwei Jiang, Qi Zhu, Tarek F. Abdelzaher, Yejin Choi, Manling Li, and Huajie Shao. ODESteer: A unified ODE-based steering framework for LLM alignment. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=CFewUmgIIL.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, Christine Moung-Wen, Brian Piening, Carlo Bifulco, Mu Wei, Hoifung Poon, and Sheng Wang. BiomedParse: A biomedical foundation model for image parsing of everything everywhere all at once. *arXiv preprint arXiv:2405.12971*, 2024. URL https://arxiv.org/abs/2405.12971.

A Datasets, Medical Concepts, and Data Splits

Our analyses require both concept-presence labels and spatial annotations, so that we can test whether a concept remains represented and how its visual evidence affects the answer. We construct image–concept questions from segmentation datasets assembled through BiomedParse (Zhao et al., 2024). Tables 5 and 6 summarize the available pools before splitting; experiment-specific cohorts can be smaller because of sampling or region eligibility.

A.1 Dataset sources and their roles

CT and MRI. AMOS (Ji et al., 2022) provides abdominal-organ annotations. We do patient based sampling from the pool for abdominal concepts. KiTS23 (KiTS23 Challenge Organizers, 2023) supplies kidney and kidney-tumor examples, and LIDC-IDRI (Armato et al., 2011) supplies lung nodules. Cardiac MRI examples come from ACDC (Bernard et al., 2018) and M&Ms (Campello et al., 2021). The LGG/TCGA subset (Buda et al., 2019) contains 257 positive brain-tumor records for region-based analyses.

Chest X-ray (CXR). The CXR pools combine COVID-QU-Ex[1], QaTa-COV19[2], the COVID-19 Radiography Database[3], and Chest Xray Masks and Labels[4] for lung, infection, and opacity concepts. SIIM-ACR (Filice et al., 2020) supplies pneumothorax annotations.

Additional modalities. The main steering comparison also includes fundus images from REFUGE (Orlando et al., 2020) and G1020 (Bajwa et al., 2020), retinal OCT from OCT-CME[5], dermoscopy from ISIC (Codella et al., 2019), pathology from PanNuke (Gamper et al., 2019), and endoscopy from PolypGen (Ali et al., 2023) and NeoPolyp[6]. Evaluating different concepts across these modalities tests the method's applicability.

External evaluation. SLAKE (Liu et al., 2021) provides the CT and MRI questions used for target-adapted transfer. IU-Xray (Demner-Fushman et al., 2016) supplies the report-generation cohort described in Appendix E.

A.2 How should the image and patient counts be read?

In Table 5 and 6, "*img*" denotes image–concept records and "*pat*" denotes patients contributing those records. Counts are per concept, including when several concepts share a row. Positive and negative refer to whether the target is visible in the supplied image or slice, not to a patient's diagnosis. A patient can therefore contribute to both columns. For example, all 300 AMOS-CT patients contribute liver-positive slices and liver-negative slices, giving 300 positive and 300 negative patients but only 300 distinct patients in total. A liver-negative slice simply lies outside the annotated liver region; it does not indicate that the patient lacks a liver. All slices from a patient remain in the same fold.

Where the export does not establish patient identity, its grouping key can instead identify an image or study. Such counts are marked with † and must not be interpreted as verified numbers of distinct patients. Source names identify the positive examples and negative examples come from the same datasets based on the annotation results.

KiTS23 sampling budget. The ordinary kidney and kidney-tumor pools each contain 600 positive and 600 negative records. The separate kidney-tumor hard-negative source pool contains 3,573 records per class from 485 patients; its negative slices contain a kidney but no annotated tumor. The configured budget is approximately 600 records *per class*, rather than 600 in total. The loader shuffles patient groups with seed 0 and retains complete groups until both classes reach this budget. Applying this rule to the current source file gives 600 positive and 611 negative records from the same 83 patients. Table 5 reports this capped pool, rather than the larger source inventory. The small overshoot preserves complete patient groups; these totals are not test-set counts.

[1] https://www.kaggle.com/datasets/anasmohammedtahir/covidqu
[2] https://www.kaggle.com/datasets/aysendegerli/qatacov19-dataset
[3] https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database
[4] https://www.kaggle.com/datasets/nikhilpandey360/chest-xray-masks-and-labels
[5] https://www.kaggle.com/datasets/zeeshanahmed13/intraretinal-cystoid-fluid
[6] https://www.kaggle.com/competitions/bkai-igh-neopolyp/data

Table 5: CT and MRI concept pools before data splitting. Pos/Neg denote target presence (Positive)/absence (Negative) in the image; img denotes image–concept records, and pat denotes contributing patients. **Counts are per concept.** The patient columns overlap and should not be added: 300 positive and 300 negative AMOS-CT patients can be the same 300 people. † marks exported image/study groups without verified patient linkage. The KiTS23 hard-negative row uses the configured cap; other rows describe available pools.

Modality	Source	Concept	Pos img	Neg img	Pos pat	Neg pat
CT — AMOS (300 patients)						
CT	AMOS-CT	liver / kidney / pancreas / aorta / postcava	600	600	300	300
CT	AMOS-CT	spleen	600	600	298	300
CT	AMOS-CT	stomach	600	600	297	300
CT	AMOS-CT	duodenum / esophagus	600	600	299	300
CT	AMOS-CT	bladder	600	600	294	300
CT	AMOS-CT	gallbladder	600	600	282	300
CT	AMOS-CT	adrenal_gland	600	600	300	300
CT — KiTS23 and LIDC-IDRI						
CT	KiTS23	kidney_tumor	600	600	98	305
CT	KiTS23	kidney_tumor (hard neg., capped)	600	611	83	83
CT	KiTS23	kidney	600	600	148	244
CT	LIDC-IDRI	nodule	600	600	159	148
MRI — AMOS (33 patients), cardiac sources, and LGG						
MRI	AMOS-MRI	kidney / liver / pancreas / stomach / duodenum / esophagus	600	600	33	33
MRI	AMOS-MRI	spleen	600	600	32	33
MRI	AMOS-MRI	adrenal_gland	480	480	33	33
MRI	AMOS-MRI	gallbladder	379	379	29	33
MRI	AMOS-MRI	postcava	146	146	33	27
MRI	ACDC / M&Ms	heart_ventricle	554	554	89	55
MRI	ACDC / M&Ms	myocardium	499	499	89	12
MRI	LGG/TCGA	brain_tumor (positive only)	257	0	257†	0

A.3 Data splits

The default protocol assigns patient groups to training, validation, and test folds in a $60/20/20$ ratio. The split uses source-specific patient identifiers where available and the exported grouping key otherwise. Balancing a complete pool also does not guarantee exact class balance within each fold.

B Concept Probing and Connector Preservation

We use complementary analyses to distinguish concept recoverability from preservation of the visual representation. Linear probes measure whether concept presence can be decoded, while similarity and reconstruction analyses examine what changes across the connector.

B.1 Linear probes and fitting partitions

Encoder and connector representations are mean-pooled visual tokens; decoder representations are residual states at the final prompt position, as defined in Section 4. A separate linear presence readout is fitted for each concept and stage while all model weights remain frozen. The probe evaluation covers all included modalities. At each visual stage, we standardize features using the fitting-set mean and standard deviation, then fit a logistic readout with at most $2{,}000$ iterations. We use the training set for fitting and evaluate on the fixed test fold.

Table 6: Additional concept pools before data splitting. img denotes image–concept records; pat denotes patients, with † marking exported groups without verified patient linkage. Positive and negative groups may overlap. Source names identify positive examples; negatives may come from other sources in the same pool. Other conventions follow Table 5.

Modality	Source	Concept	Pos img	Neg img	Pos pat	Neg pat
Chest X-ray						
CXR	COVID-QU / Radiog. / CXR Masks	lung	600	600	600†	600†
CXR	Radiography	lung_opacity	600	600	600†	600†
CXR	COVID-QU / QaTa	covid-19_infection	600	600	600†	600†
CXR	SIIM	pneumothorax	290	290	290†	290†
Ultrasound						
US	CAMUS	heart_ventricle / heart_atrium	600	600	100	600†
US	FH-PS-AOP	fetal_head	600	600	600†	228†
US	BreastUS	tumor	128	128	128†	128†
Additional modalities						
Other	REFUGE / G1020	optic_disc	600	600	600†	600†
Other	REFUGE / G1020	optic_cup	558	558	558†	558†
Other	OCT-CME	edema	283	283	283†	283†
Other	ISIC	lesion	600	600	600†	600†
Other	PanNuke	neoplastic_cells	600	600	600†	600†
Other	PolypGen / NeoPolyp	polyp	499	499	499†	499†

B.2 Representational similarity and reconstruction

Representational transformation. To test whether the connector preserves visual feature structure, we compare paired encoder and connector representations for each concept. For SVCCA (Raghu et al., 2017), we center each feature matrix, retain the leading 20 singular directions (or all available directions if fewer), and average the canonical correlations between the reduced representations.

Following the notations in our analysis, we here denote the pooled visual representations produced by the encoder and connector as $\mathbf{h}_{-1}(x) \in \mathbb{R}^{d_{-1}}$ and $\mathbf{h}_0(x) \in \mathbb{R}^{d_0}$, respectively. For reconstruction, we fit a ridge regression map with fixed penalty 1 from pooled connector features $\mathbf{h}_0(x)$ to pooled encoder features $\mathbf{h}_{-1}(x)$. The connector features are standardized using statistics computed only on the fitting partition. The reported test score is

$$R^2(0 \to -1) = 1 - \frac{\sum_{i \in \text{test}} \left\| \mathbf{h}_{-1}(x_i) - \widehat{\mathbf{h}}_{-1}(x_i) \right\|_2^2}{\sum_{i \in \text{test}} \left\| \mathbf{h}_{-1}(x_i) - \bar{\mathbf{h}}_{-1,\text{test}} \right\|_2^2}, \tag{4}$$

where $\widehat{\mathbf{h}}_{-1}(x_i)$ is the encoder representation predicted from $\mathbf{h}_0(x_i)$, and

$$\bar{\mathbf{h}}_{-1,\text{test}} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{i \in \text{test}} \mathbf{h}_{-1}(x_i)$$

is the mean encoder representation over the test fold.

This score measures the fraction of variance in the pooled encoder representations that can be recovered by a linear map from the pooled connector representations. We report this connector-to-encoder reconstruction in Table 1; a high score indicates that the pooled encoder representation remains largely linearly recoverable after the connector. Because reconstruction is directional, the reverse mapping $R^2(-1 \to 0)$ can yield a different score. Importantly, reconstruction of pooled representations does not establish preservation of every concept or fine-grained target region.

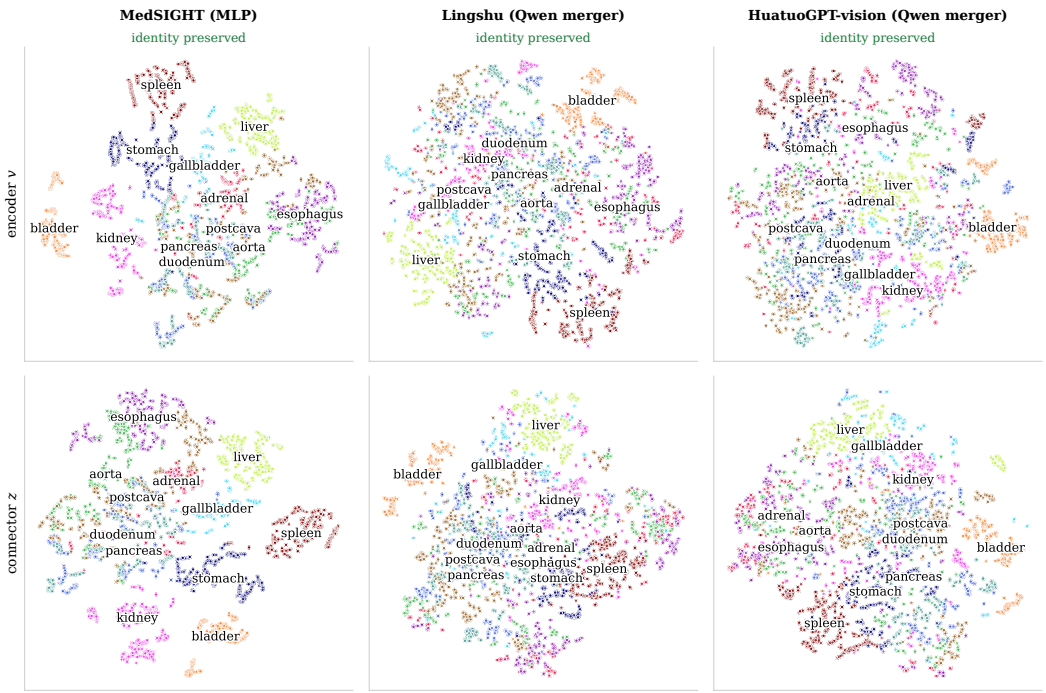


Figure 7: Organ-region feature organization before and after the connector. Columns show MedSIGHT, Lingshu, and HuatuoGPT-Vision; rows show encoder (top) and connector (bottom) features. Colors indicate organ identity, with matched region samples across stages within each model. Each panel is fitted independently; visible grouping provides qualitative evidence of organ-associated structure, not a test of geometric or spatial preservation.

B.3 Concept-level visualization before and after the connector

Figure 7 compares t-SNE projections of AMOS organ-region features before and after the connector for MedSIGHT, Lingshu, and HuatuoGPT-Vision. Unlike the image-level pooled representations analyzed above, each point represents a feature vector pooled over an annotated organ region. Within each model, the same sampled regions are used at both stages, with at most 350 examples per organ and consistent concept colors across panels. Features are standardized and reduced to at most 50 principal components before fitting t-SNE separately for each model and stage, using perplexity 30, 1,000 iterations, and random seed 0.

Organ-associated grouping is visible at both stages, although separation varies across concepts and models. MedSIGHT shows relatively distinct groups for organs such as the spleen, bladder, and kidney in both encoder and connector features. Lingshu and HuatuoGPT-Vision exhibit more overlap among organ groups, while several concepts, including the bladder and spleen, occupy recognizable concentrations at both stages. These patterns qualitatively support the presence of organ-associated structure after projection, without implying uniform separation of all concepts from the pretrained encoders themselves.

B.4 Probe Data Efficiency

Throughout the paper, the presence probe is fit on the full training fold, and the same frozen probe supplies the steering gate (Section 4, Section 7.1). This raises a fair concern: perhaps concept decodability, and any downstream steering leverage, is an artifact of large-scale probe supervision rather than a property of the representation. If the probe needed thousands of labels to reach its reported accuracy, "the concept is decodable" would be closer to "we trained a strong classifier." We therefore measure how the *same* probe behaves as its training pool is shrunk.

Table 7: MedSIGHT presence-probe balanced accuracy versus per-class training size ($\ell^\star = 27$), mean over 5 draws, averaged within each modality and then across modalities (equal weight). "Full" uses the entire training pool. Even at 8 labels/class the modality average is $0.88\times$ the full-data value, reaching $0.94\times$ by 64 and $0.99\times$ by 256.

Modality	n=8	n=16	n=32	n=64	n=128	n=256	Full	#concepts
CT (incl. findings)	0.861	0.875	0.885	0.890	0.897	0.933	0.933	7
MRI (organs)	0.728	0.758	0.779	0.795	0.807	0.822	0.837	10
CXR (findings)	0.761	0.760	0.787	0.826	0.823	0.884	0.890	4
Average	**0.783**	**0.798**	**0.817**	**0.837**	**0.842**	**0.879**	**0.887**	21

Protocol. We take the decision-layer residual representation $\mathbf{h}_{\ell^\star}(x, q(c))$ at the final prompt position, i.e. the exact tap used by the layer-wise probe and the steering gate ($\ell^\star = 27$ for MedSIGHT). Holding the fixed seed-0 test fold fixed, we draw balanced subsamples of $n \in \{8, 16, 32, 64, 128, 256\}$ examples per class from the training pool, refit the identical linear probe on each subsample, and report balanced accuracy on the full test fold, averaged over 5 independent draws. This analysis reuses the exact representations, folds, probe family, and metric of the main experiments. We evaluate 21 concepts across three modalities: 7 CT concepts (including the nodule and kidney-tumor findings), 10 abdominal MRI organs, and 4 chest X-ray findings. We summarize by averaging concepts within a modality and then averaging the three modalities with equal weight, so no modality dominates by concept count.

Findings. Averaged across the 21 concepts, balanced accuracy rises modestly with training size and is already high at the smallest budget (Figure 8, Table 7). With only 8 labels per class the probe reaches 0.783, which is 88% of its full-training value (0.887); by 64 labels per class it recovers 94%, and 95% by 128. Gains beyond a few dozen labels are small (a further $\sim$0.05 from n=64 to the full $\approx$ 950/class pool), i.e. the vast majority of the decodable signal is available from a handful of labels. The trend holds within every modality; CT saturates fastest (already 0.86 at n=8), while the MRI-organ and CXR pools start lower and keep the same shallow slope. This pattern is not specific to $\ell^\star$: at earlier and later layers ($\ell \in \{13, 20, 31\}$) the curves are near-identical (e.g. nodule reaches 95% of full at n=8 and kidney tumor at n=16–64). Because the steering gate is exactly this probe, neither decodability nor the gate depends on large-scale probe training. This complements the steering training-count ablation in Appendix D.4, which shows the intervention itself is likewise effective from few examples: together the two establish that the method's leverage is not bought with data scale on either the probe or the direction.

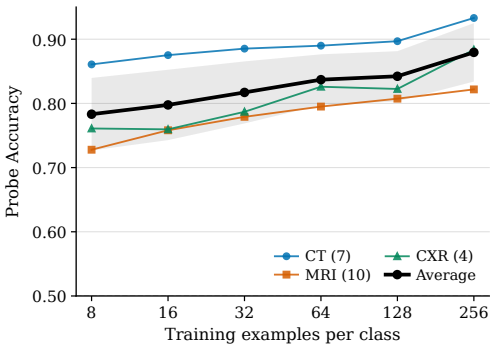


Figure 8: **Presence-probe data efficiency.** Balanced accuracy of the MedSIGHT presence probe at the decision layer ($\ell^\star = 27$) versus the number of training examples per class. The bold line shows the average across modalities, and the dashed line denotes chance performance.

C Concept Use Within the Decoder

We examine why recoverable concept information can still lead to incorrect answers. Attention and activation analyses identify differences associated with correctness, while activation patching tests whether restoring a component changes support for concept presence.

C.1 TARGET-REGION ATTENTION ACROSS THREE MODELS

Figure 9 extends the attention analysis in Section 6.1 to HuatuoGPT-Vision. All three models are evaluated on the same 80 positive CT question–image records from 63 patients: 40 nodules and 40 kidney tumors, sampled with seed 0. At the final prompt position, we sum attention over target-region visual tokens and average over heads within each layer.

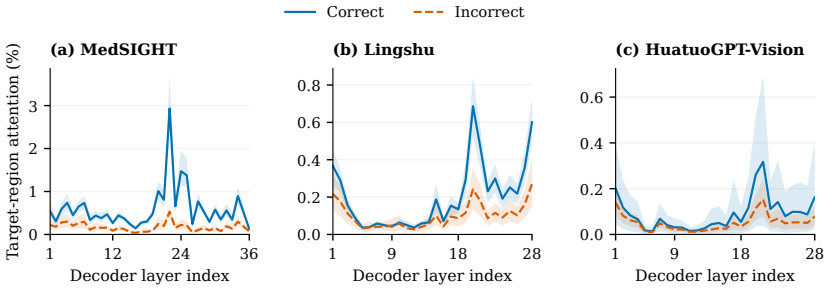


Figure 9: **Target-region attention on a matched CT cohort.** Head-averaged attention from the final prompt position, macro-averaged over nodule and kidney tumor within each correctness group. Shading shows pointwise patient-bootstrap intervals. Decoder layers are one-based; vertical scales differ between models. The association is descriptive and does not establish a causal effect of attention.

C.2 CORRECTNESS SEPARATION IN CT MLP ACTIVATIONS

To test whether correctness separation extends across models, we apply the analysis in Section 6.2 to identical CT records for MedSIGHT, Lingshu, and HuatuoGPT-Vision. We use a common fitting and evaluation protocol for all three models.

For each concept and layer, we subtract the concept's fitting-set mean MLP output from both fitting and held-out activations. We fit an unstandardized logistic correctness classifier ($C = 1$, maximum 2,000 iterations) per layer. LDA is fitted on the same fitting partition and displayed at each model's observed peak held-out AUROC. This layer choice is descriptive, not an independently validated selection rule. AUROC uses all held-out records; histogram scores are clipped to their pooled 1st/99th percentiles for display only, using 34 common bin edges within each model. All three models use the same fitting, centering, and plotting protocols. Figure 10 shows the layerwise correctness AUROC and the held-out LDA distributions for all three models; Table 8 summarizes their observed peaks.

Table 8: CT correctness separation with a shared patient partition. Peak layers are one-based and selected descriptively from the displayed curve.

Model	Peak layer	AUROC	Held-out correct	Held-out incorrect
MedSIGHT	27	0.850	225	160
Lingshu	20	0.909	281	104
HuatuoGPT-Vision	22	0.890	258	127

C.3 ACTIVATION PATCHING ACROSS MODELS AND MODALITIES

Figure 11 extends the CT analysis to MedSIGHT, Lingshu, and HuatuoGPT-Vision across the evaluated concept groups. At each tested layer, we patch the attention or MLP output separately and average validation-set recovery scores equally across concepts within each group. CT pools liver, spleen, kidney, nodule, and kidney tumor; MRI contains the three abdominal organs; CXR contains lung and pneumothorax; and US contains fetal head, heart atrium, heart ventricle, and tumor. Brain-MRI reports brain tumor separately. The row labeled OOD in the figure pools concepts from other modalities such as edema, lesion, neoplastic cells, optic cup, optic disc, and polyp.

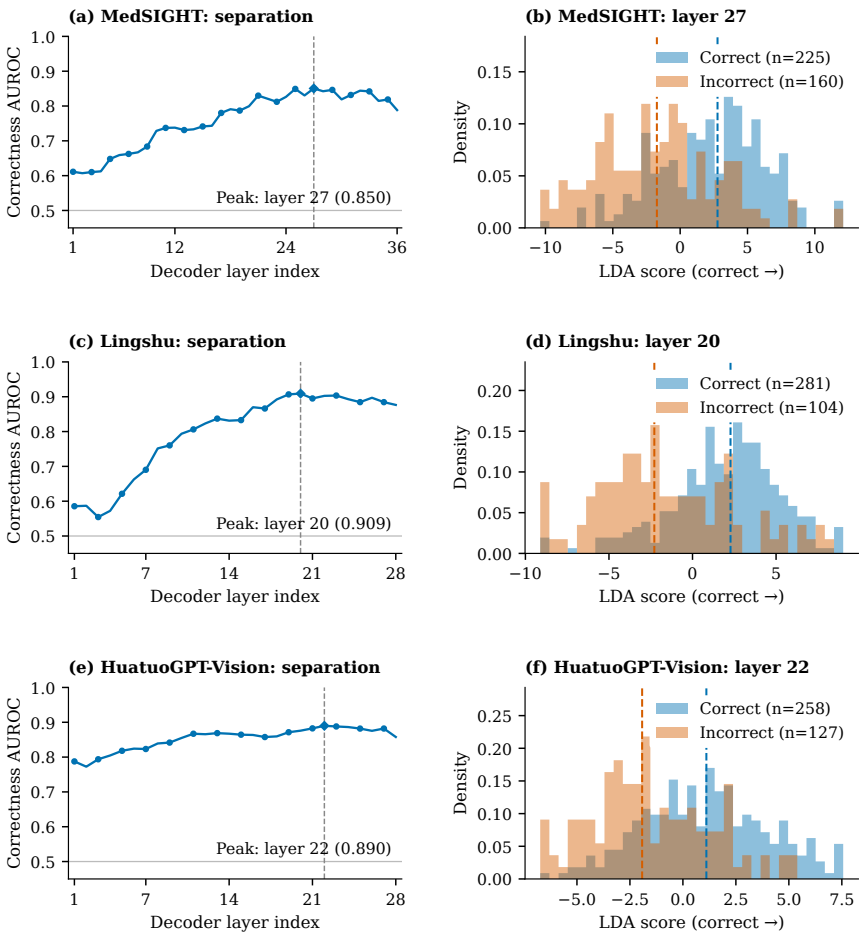
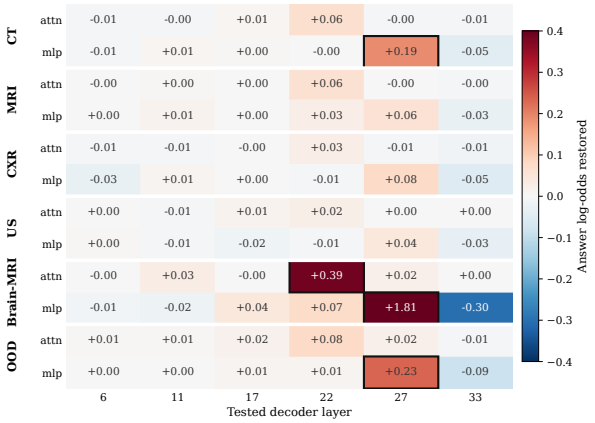


Figure 10: **CT activation contrasts in a common format.** Rows show MedSIGHT, Lingshu, and HuatuoGPT-Vision on identical records and patient partitions. Left: held-out correctness AUROC at every decoder layer; the dashed vertical line marks the observed peak. Right: held-out LDA scores at that layer, with dashed group medians. Histogram clipping affects display only. Lingshu reproduces the main-paper analysis; the other two models use the same protocol.

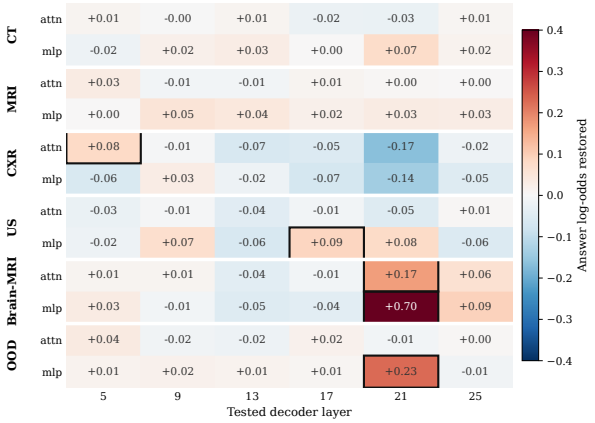
Where does restoration have the largest effect? MedSIGHT shows its largest MLP recovery at index 27 across all six groups, although the effect varies substantially: brain tumor reaches 1.81 log-odds, compared with 0.19 for CT and 0.04 for US. Lingshu also peaks at MLP index 21 for CT, Brain-MRI, and OOD, but its MRI and US MLP peaks occur earlier, at indices 8 and 16. Its CXR MLP effects are mostly negative, and the largest positive CXR effect comes from attention at index 4 (0.08). HuatuoGPT-Vision instead has its strongest MLP recovery at index 17 for CT, MRI, CXR, and US, and at index 21 for Brain-MRI and OOD.

Overall, these results suggest that restoration effects are often concentrated within a relatively narrow range of decoder layers, but the exact optimal component and effect magnitude remain concept- and modality-dependent. To avoid introducing a separate intervention site for every concept or modality, we therefore use a single model-level steering layer in main experiments, selected by aggregating restoration effects across groups. This provides a shared intervention site that captures the model's overall region of strongest causal recovery while keeping the steering procedure simple and scalable.

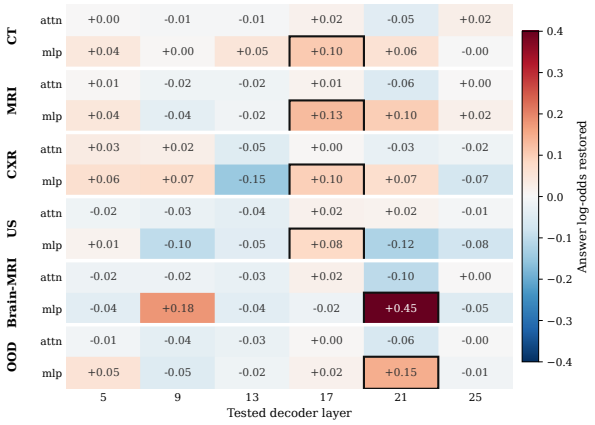
Why do recovery effects differ across groups? Differences in training exposure to imaging modalities may contribute to these patterns, as may the type of concept being queried. For example, the abdominal MRI group tests anatomical structures, whereas Brain-MRI tests a focal abnormality; the much larger Brain-MRI effects therefore cannot be attributed to imaging modality alone. Mixed groups such as CT and US also combine anatomy and abnormalities, whose spatial extent



(a) MedSIGHT



(b) Lingshu



(c) HuatuoGPT-Vision

Figure 11: **Activation patching across models and modalities.** Cells show concept-averaged changes in presence-answer log-odds after restoring one clean attention or MLP output in the target-masked run. Positive values indicate recovery; negative values indicate reduced support. Boxes mark each row's maximum when it exceeds 0.08. The shared color scale is capped at ± 0.40; annotations retain actual values beyond this range.

and diagnostic cues can differ. These are possible explanations rather than factors isolated by this experiment. Mask size, the initial masking effect, and answer-score calibration can also influence the raw recovery scores. The heatmaps support model- and concept-dependent localization of recoverable evidence, without establishing that a larger score reflects better recognition or stronger clinical performance.

D STEERING IMPLEMENTATION AND EVALUATION

This section explains how we fit and evaluate MEDSTEER, then examines the roles of the gate, direction, intervention depth, and training budget. The analyses test when concept-guided edits help turn recoverable information into correct answers.

D.1 MED-LVLMS

We evaluate the publicly released checkpoints below, keeping all model weights frozen during probing and activation steering. Lingshu, MedSIGHT, and HuatuoGPT-Vision are used in the main steering comparison.

Lingshu. We use Lingshu-7B (Xu et al., 2026), a medical vision–language model based on Qwen2.5-VL. Its visual encoder and patch-merging connector provide a complementary interface to MedSIGHT's CLIP-based encoder and MLP projector.

MedSIGHT. We use MedSIGHT-Comp, the comprehension-only variant of MedSIGHT (Chang et al., 2026). This variant removes the segmentation branch and retains the visual encoder, vision–language connector, and language decoder for medical visual comprehension. It combines a UniMed-CLIP encoder, an MLP projector, and a Qwen3-8B language model. Throughout this paper, *MedSIGHT* refers to this MedSIGHT-Comp checkpoint. The official repository describes the full model and its comprehension-only variant.

HuatuoGPT-Vision. We use HuatuoGPT-Vision-7B-Qwen2.5VL (Chen et al., 2024), the Qwen2.5-VL variant of HuatuoGPT-Vision. We refer to this checkpoint as *HuatuoGPT-Vision* in the three-model steering and decoder analyses.

D.2 TRAINING AND HYPERPARAMETER SELECTION

For each concept, we fit the presence probe and estimate the original-minus-masked activation direction using training examples. The probe score r_c is a logistic decision score, so a threshold of $\tau_c = 0$ corresponds to predicted presence probability 0.5.

Threshold and strength selection. We select the concept-specific threshold τ_c by grid search on the validation set over

$$\tau_c \in \{-0.5, -0.25, 0, 0.25, 0.5\}\, \sigma_{c,\mathrm{val}},$$

where $\sigma_{c,\mathrm{val}}$ denotes the standard deviation of the probe score on the validation set. Empirically, the selected threshold is $\tau_c = 0$ for most concepts across all three models, corresponding to the probe's standard decision boundary at a predicted presence probability of 0.5. For each concept, we select $\alpha_c \in \{0, 1, 2, 4, 8, 16, 32\}$ using up to 24 positive and 24 negative validation examples, choosing the value that maximizes the model's edited-answer balanced accuracy. Here $\alpha_c = 0$ denotes the unedited model. The probe, direction, and selected settings remain fixed during test evaluation. The threshold can be recalibrated on validation data for other datasets or settings. Steering-strength (α_c) ablations and results are presented in Appendix D.4.

Balanced accuracy by modality. Table 2 reports the modality comparison directly. For each model and eligible concept–modality pair, balanced accuracy is the mean of sensitivity and specificity. We then average over the modality level results including CT, MRI, CXR, and Others. Others comprises pathology, fundus, OCT, dermoscopy, and endoscopy.

D.3 GATE, DIRECTION, AND DEPTH ANALYSIS ACROSS THREE MODELS

To test whether the gate, direction, and depth findings extend to a third model, we repeat the analysis in Figure 5 for HuatuoGPT-Vision and compare all three models on a shared CT test sample.

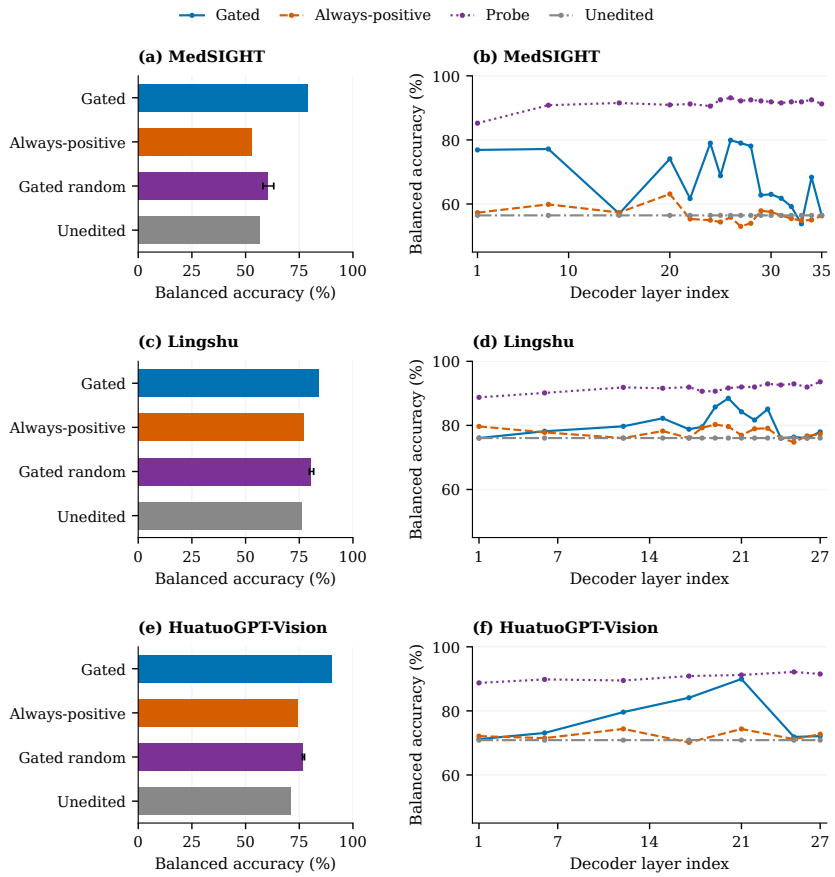


Figure 12: **Gate, direction, and depth analysis on a shared CT test sample.** Scores are macro balanced accuracy over five concepts. (a,c,e): gated steering, always-positive steering, gated random edits, and unedited answers at the reference site (MedSIGHT layer 27; Lingshu and HuatuoGPT-Vision layer 21). (b,d,f): steering and probe accuracy across measured depths. HuatuoGPT-Vision uses seven predeclared depths; the other models are rescored from saved predictions. Lines connect measured points without imputing missing layers.

Figure 12 shows that gated steering outperforms both always-positive steering and gated random edits at the reference sites. The depth curves further show that strong concept decodability does not guarantee effective steering at every layer.

Evaluation protocol. We use the existing 1,234 CT test records, stratified by concept and presence label. Every model and intervention setting uses these same records. We report the equal-weight mean of the five concept balanced accuracies.

At each intervention layer, the presence probe is fitted on the full training fold, and the raw intact-minus-target-masked MLP direction is estimated from 64 positive training candidates per concept. Reference sites are layer 27 for MedSIGHT and layer 21 for Lingshu and HuatuoGPT-Vision. Random-direction controls preserve the learned direction's norm and the gate, and each of seeds $0, 1, 2$ is independently tuned on the same validation grid.

Contributions of the gate and learned direction. Figure 12(a,c,e) shows that gated steering achieves 79.0%, 84.3%, and 89.9% macro balanced accuracy for MedSIGHT, Lingshu, and HuatuoGPT-Vision, respectively, exceeding both controls in each model. For HuatuoGPT-Vision, the corresponding accuracies are 70.9% unedited, 74.4% with always-positive steering, and 76.8% with gated random directions. Gated steering therefore improves on these conditions by 19.1, 15.6, and 13.1 percentage points, respectively (computed before rounding). These independently tuned

comparisons support the contributions of both input-dependent gating and the learned concept direction.

Intervention depth and concept decodability. For HuatuoGPT-Vision, we evaluate seven predeclared layers, $1, 6, 12, 17, 21, 25, 27$, to cover early, middle, and late depths while limiting computation. Each depth has its own training-fitted probe and direction; MedSIGHT and Lingshu retain their previously measured depths. The final decoder layer is excluded. Figure 12(f) shows HuatuoGPT-Vision's gated accuracy increasing from 71.2% at layer 1 to 84.1% at layer 17, with the highest measured value of 89.9% at the existing layer-21 reference site. Accuracy then falls to 71.9% and 72.1% at layers 25 and 27, close to the 70.9% unedited baseline, although probe accuracy remains between 88.7% and 92.2% across the measured layers. MedSIGHT and Lingshu likewise show substantial depth variation in steering effectiveness despite strong probe performance. Thus, decodable concept information alone does not identify an effective site for this intervention. These are exploratory results on a sample of the existing patient-disjoint test fold; the curves connect measured points and do not establish an optimum at unmeasured layers.

D.4 Training-example count and steering-strength ablations

We next ask how many training examples steering needs and how sensitive it is to edit strength, extending Section 7.3. We use the same five CT concepts and MLP sites as Figure 5(a): layer 27 for MedSIGHT and layer 21 for Lingshu. The sampled records are identical across models and settings and we use the existing patient-disjoint training, validation, and test folds. We report the equal-weight mean of the five concept balanced accuracies.

Training-example count ablation. We vary $N \in \{8, 16, 32, 64\}$ training examples *per class*, using nested subsets and refitting both the presence probe and the direction at each budget. At each budget, we select α_c separately for each concept from $\{0, 1, 2, 4, 8, 16, 32\}$ on a fixed sample of 8 validation examples per class, maximizing balanced accuracy and breaking ties toward smaller strengths. The selections are frozen before test scoring.

Steering-strength ablation. We fix the probe fitted on the full training fold and the original direction estimated from 64 positive training candidates, with the same token-overlap exclusions. We then apply each absolute strength $\alpha \in \{0, 1, 2, 4, 8, 16, 32\}$ to all five concepts, keeping the probe, direction, and intervention site fixed. Here $\alpha = 0$ is the unedited model; the remaining points measure sensitivity to edit strength and are not used to select a test-optimal setting. Unlike the count sweep, this sweep retains all training-fold examples for the probe.

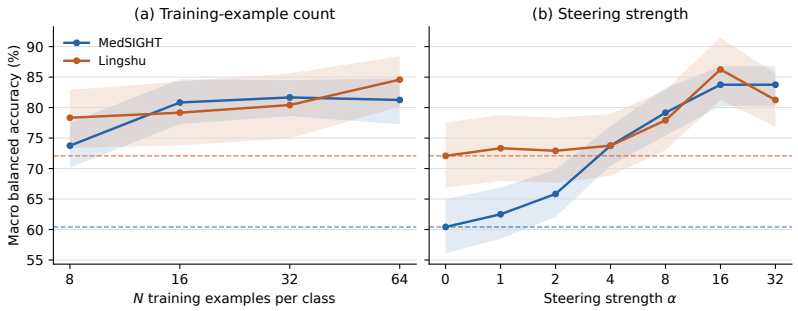


Figure 13: Additional CT ablations on a fixed, class-balanced test sample. (a) Varying direction training budgets, with strength selected on validation data at each budget. (b) Varying the absolute edit strength with the original direction fixed. Dashed lines show unedited accuracy on the same sample. Shaded regions are 95% percentile intervals from $2{,}000$ patient-bootstrap replicates, resampling jointly across concepts and settings. Intervals are conditional on the fitted probes, directions, validation selections, and fixed test sample; they do not include training-subsample variability.

Results. Figure 13 shows that increasing the training budget from $N = 8$ to $N = 64$ raises balanced accuracy from 73.8% to 81.2% for MedSIGHT and from 78.3% to 84.6% for Lingshu.

MedSIGHT's gains largely plateau by $N = 16$, whereas Lingshu continues to improve through $N = 64$ in this sampled sweep. Strength sensitivity differs between models: MedSIGHT reaches 83.8% at both $\alpha = 16$ and 32, while Lingshu drops from 86.2% at $\alpha = 16$ to 81.2% at 32. Thus, increasing the edit magnitude does not uniformly improve recognition. These strength comparisons are descriptive and do not replace validation-based hyperparameter selection.

D.5 Baseline implementation details

To compare interventions under a common evaluation protocol, the baselines share the data, $60/20/20$ group folds, concept set, validation-based selection, and full test pool. Patient grouping follows Appendix A.3. The comparison covers MedSIGHT, Lingshu, and HuatuoGPT-Vision. Adaptations and differences in fitting resources are described below.

CAA (contrastive activation addition). CAA (Rimsky et al., 2024) steers with a single *contrastive activation* direction. As a medical adaptation, we form this direction from the label contrast rather than from paired prompts: at the shared intervention site $(\ell^\star, m^\star)$, we take the class-mean difference of the residual-stream state at the answer position $p^\star$ between concept-present and concept-absent training examples,

$$\mathbf{g}_c^{\mathrm{CAA}} = \mathbb{E}[\mathbf{h}_{\ell^\star} \mid c \text{ present}] - \mathbb{E}[\mathbf{h}_{\ell^\star} \mid c \text{ absent}],$$

estimated from cached clean activations over the training fold. This is distinct from our direction $\mathbf{g}_c$ (Eq. 2), which contrasts the *same* image with and without its target region at the MLP output; the CAA direction instead contrasts *different* images by label and, following the original method, is added to the residual stream (the residual-stream intervention used by CAA).

CAA applies a single *fixed* edit: one signed multiplier α_c is added to every test example's residual state, with no per-example presence gate. This is the key difference from our gated update (Eq. 3), which uses the probe to select the edit sign or abstain. We select α_c per concept on the validation fold over a small signed grid that includes $\alpha_c{=}0$ (the unedited model), maximizing balanced accuracy; the chosen edit is then evaluated on the untouched test fold. The sign of the CAA edit is fixed within each concept, whereas our gate selects the sign separately for each input. For CAA, we use the full training set rather than only 64 examples.

VTI (visual and textual intervention). VTI (Liu et al., 2025) is a test-time shift that stabilizes vision features by steering hidden states toward the representation of a clean image, using a direction obtained by contrasting a clean image against heavily corrupted copies. As a medical adaptation we fit this direction *per concept* rather than as a single dataset-level vector: for each of a concept's training images we contrast its clean answer-position hidden states with those of $T{=}4$ patch-masked copies (a fraction 0.99 of image patches set to the image mean), average over copies, and take the rank-1 principal direction of the clean$-$corrupted differences at every decoder layer. At inference this per-layer direction is added at all token positions to each decoder-layer MLP output as a normalized shift, renormalized to preserve the hidden-state norm, with a single per-concept strength α_c (including $\alpha_c{=}0$) selected on the validation fold; like CAA, the edit is applied to all examples without a presence gate. We implement VTI's *textual* (decoder) branch only. Its visual-encoder branch is tied to the vision-tower architecture (CLIP for MedSIGHT versus the Qwen2.5-VL resampler for Lingshu and HuatuoGPT-Vision) and is omitted, following the paper's own ablation that attributes most of the effect to the decoder branch.

V-ITI (visual inference-time intervention). V-ITI (Sun et al., 2025) is a head-level attention edit that, at each decoder layer, blends every attention head's output toward its *visual-recall* vector: the same head's attention output but renormalized to attend only to the image tokens. For head (ℓ, h) we reconstruct the value-weighted output at the answer position, $\mathbf{o}_\ell^h = \mathbf{w}_\ell^h \mathbf{V}_\ell^h$, from the eager attention weights and value projection (exact, since rotary and query/key normalization do not act on values) and its image-token counterpart $\boldsymbol{\mu}_\ell^h$ over the resolved visual span, and apply $\hat{\mathbf{o}}_\ell^h = (1{-}\alpha)\mathbf{o}_\ell^h + \alpha\,\boldsymbol{\mu}_\ell^h$ re-injected through the output projection. Intervention is head-selective and *when*-gated by a per-head discriminative probe $p_\theta(\mathbf{o}_\ell^h) = \sigma\langle\theta, \mathbf{o}_\ell^h\rangle$ with $\alpha = \alpha_0\, p_\theta$. As a medical adaptation, since we have no hallucination annotations, we fit each probe on the training fold with the neglect label set to the model's own unedited error, pooling across a column's concepts (V-ITI forms no concept-specific direction); we keep the paper's top-10% head selection and select the single global strength

α_0 on the validation fold. Unlike our method, V-ITI defines its own multi-layer attention locus rather than the site $(\ell^\star, m^\star)$, and requires a resolved image-token span, available for the three Qwen-family backbones.

VISTA (visual information steering). VISTA (Li et al., 2025) reinforces visual evidence during decoding by adding a *visual steering vector* (VSV) to the residual stream. Its defining choice is the contrast used to build the direction: for each demonstration it runs the same text prompt *with* and *without* the image and takes the paired difference of answer-position hidden states, isolating the residual-stream shift that carries visual information—in contrast to VTI, which perturbs image pixels rather than removing the image. We take the rank-1 principal direction of these with−without differences at every decoder layer and, as a medical adaptation, fit it *per concept* from that concept's training images. At inference the per-layer direction is added at all token positions to the decoder-layer MLP output as a normalized shift, renormalized to preserve the hidden-state norm, with a single per-concept strength α_c (including $\alpha_c{=}0$) selected on the validation fold.

ODESteer (ODE-based steering). ODESteer (Zhao et al., 2026) replaces one-step vector addition with a *multi-step, per-example-adaptive* edit. Its defining object is a *barrier function* $h(\mathbf{a}) = \log\big(p_+(\mathbf{a})/p_-(\mathbf{a})\big) = \mathbf{w}^\top \phi(\mathbf{a}) + b$, the log density-ratio between concept-present and concept-absent activations, which we fit by (regularized) logistic regression on cached clean MLP output states $\mathbf{a}$ at the shared site $(\ell^\star, m^\star)$; ϕ is a degree-two polynomial random-feature map (polynomial count sketch, 8000 features, $\gamma{=}0.1$, $c_0{=}1$) that keeps the kernel differentiable. The edit integrates the steering ODE $\dot{\mathbf{a}} = \nabla_{\mathbf{a}} h(\mathbf{a})/\|\nabla_{\mathbf{a}} h(\mathbf{a})\|$ by the Euler method for ten steps, so the direction re-adapts at each step from the current activation rather than being fixed; the gradient is taken by autograd through ϕ. Because the original formulation is unidirectional (always toward the positive class) whereas presence is two-sided, we integrate along $\eta_c \nabla_{\mathbf{a}} h$ using the *same* frozen presence gate $\eta_c{=}\mathrm{sign}(r_c)$ as our method; the only difference from our update (Eq. 3) is then the field itself—a multi-step density-ratio gradient versus the single mean difference $\mathbf{g}_c$. The integration horizon T is set as a fraction of the fit-set mean activation norm (scale-invariant across backbones) and selected per concept on the validation fold, including $T{=}0$ (no edit). The barrier is fit on the full training fold rather than the 64 positives our direction uses, since a density-ratio classifier over 8000 features needs more samples than a mean vector. We compare against ODESteer's steering edit under our shared prediction rule.

Training data for steering baselines. Single-direction methods (MEDSTEER, VTI, VISTA) require only a few dozen samples for stable estimation, and we therefore use 64 training examples. Methods that fit higher-dimensional objects use the full training fold, including ODESteer's 8,000-feature density-ratio classifier and CAA's class-mean estimator.

Contrastive-decoding baselines (Med-VCD, M3ID, PAI, DoLa). The second baseline group mitigates hallucination at *decoding* time rather than by editing activations: each method modifies the answer-position next-token distribution, from which we read the same presence decision $\mathrm{sign}(\log P(\text{yes}) - \log P(\text{no}))$ as every other row. All four baselines except DoLa share the training-free form $\boldsymbol{\ell}^{\text{adj}} = (1{+}\alpha_c)\log p(y \mid v) - \alpha_c \log p(y \mid \tilde{v})$ followed by the adaptive plausibility constraint of visual contrastive decoding (retain only tokens whose conditional log-probability lies within $\log\beta$ of the maximum, $\beta{=}0.1$), and differ only in the counterfactual $\tilde{v}$ and, for PAI, an additional attention edit. Because these methods target one answer token, we use them in the constant-strength, answer-position form standard for yes/no hallucination benchmarks: the contrast strength α_c is selected per concept on the validation fold over a small grid including $\alpha_c{=}0$ (the unedited model), exactly as for our steering strength, and the counterfactual forwards are produced by perturbing the pixel input to the vision encoder so the procedure is identical across backbones. **Med-VCD** (Mahdavi et al., 2026) contrasts the image against a diffusion-noised copy of the same image ($\tilde{v}$ obtained at forward noising step 500 of the standard schedule), suppressing tokens that survive when the visual evidence is destroyed. **M3ID** (Favero et al., 2024) instead contrasts against an image-free forward ($\tilde{v}$ a zeroed image), amplifying image-conditioned over language-prior tokens; its original position-decaying schedule collapses to this constant contrast at the single answer token. **PAI** (Liu et al., 2024) additionally up-weights attention to image tokens during prefill, adding $\alpha|\cdot|$ with $\alpha{=}0.2$ to the answer row's attention logits over the visual span across layers $[2, L)$, and combines this with the same image-free contrast; the attention edit is active on the two Qwen2.5-VL backbones,

where the visual span is resolvable, and reduces to the image-free contrast on MedSIGHT, whose inline image-embedding expansion leaves no explicit image-token span. **DoLa** (Chuang et al., 2024) contrasts the model's own layers instead of the visual input: it takes $\tilde{v}$ to be a *premature* distribution from an intermediate layer (early-exited through the shared final norm and unembedding), using $\log p_{\text{mature}}(y) - \alpha_c \log p_{\text{premature}}(y)$ under the same plausibility constraint, with the premature layer chosen per example by maximal Jensen–Shannon divergence from the mature one over a strided candidate set spanning the lower two-thirds of the decoder; both distributions come from a single forward, so it too is backbone-agnostic (α_c=1 recovers canonical DoLa). We do not compare against OPERA (Huang et al., 2024): its over-trust penalty and retrospection-allocation operate over a *decoded token sequence* under beam search and have no single-token form, so on a one-token presence answer OPERA reduces to ordinary beam decoding and is not comparable under our answer-position prediction rule.

E Open-Ended Generation

To test whether concept steering helps models express findings beyond yes/no answers, we evaluate renal descriptions and IU-Xray reports (Table 4). Both models use the same examples within each task, with each steered condition compared against its corresponding unsteered baseline. We do not use any training or validation samples from the new datasets; instead, we directly transfer the concept directions obtained from the previous closed-ended VQA tasks.

E.1 Renal descriptions

The KiTS23 (KiTS23 Challenge Organizers, 2023) CT cohort contains 100 test patients, all excluded from the source CT cohort. Each patient contributes one tumor-present and one kidney-present, tumor-absent slice, giving 200 test examples. The prompt is: *"Describe the kidneys in this CT image in one sentence, including any visible focal abnormality. Answer in English."* We retain the source kidney-tumor direction at decoder layer 27 for MedSIGHT and 21 for Lingshu (one-based). Decoding is greedy with a 64-token cap.

We score whether the generated description correctly asserts the target finding, using explicit tumor, mass, neoplasm, or carcinoma terminology and accounting for local negation and uncertainty. We report balanced accuracy across target-present and target-absent examples; generic lesion and cyst mentions do not count. Invalid outputs count as incorrect. This metric measures target terminology rather than complete report quality; 13 tumor-negative test slices contain annotated kidney cysts. Scoring rules were revised after inspecting test text, then applied uniformly to validation candidates for reselection, so this evaluation is exploratory.

E.2 IU-Xray reports

The local IU-Xray (Demner-Fushman et al., 2016) annotation contains 2,069 training, 296 validation, and 590 test studies. Both available views are supplied together. To evaluate our target concept "lung opacity", we sample 100 test studies for both models: 30 lung-opacity-positive and 70 non-positive references. Eligible reports contain at least one positive abnormal-finding label. Studies are ordered by a fixed hash, without individual case substitutions.

CheXbert (Smit et al., 2020) labels reference and generated reports. Target balanced accuracy measures lung-opacity agreement; positive-label micro F1 covers all 14 findings or the 13 non-target findings. Blank or negative reference labels are nonpositive, and uncertain references are excluded. Invalid outputs or uncertain target predictions count as incorrect for target accuracy; invalid text contributes no positive findings to F1. These metrics measure agreement with reference-report findings; they do not provide independent clinical assessment of the images.

E.3 Report-generation settings

Both models reuse frozen source CXR intact-minus-lung-region-masked MLP directions at decoder layers 27 and 21, respectively; no IU-Xray example is used to estimate a direction. Lingshu uses the prompt *"Write a concise report of the visible findings on this chest X-ray in one sentence. Answer*

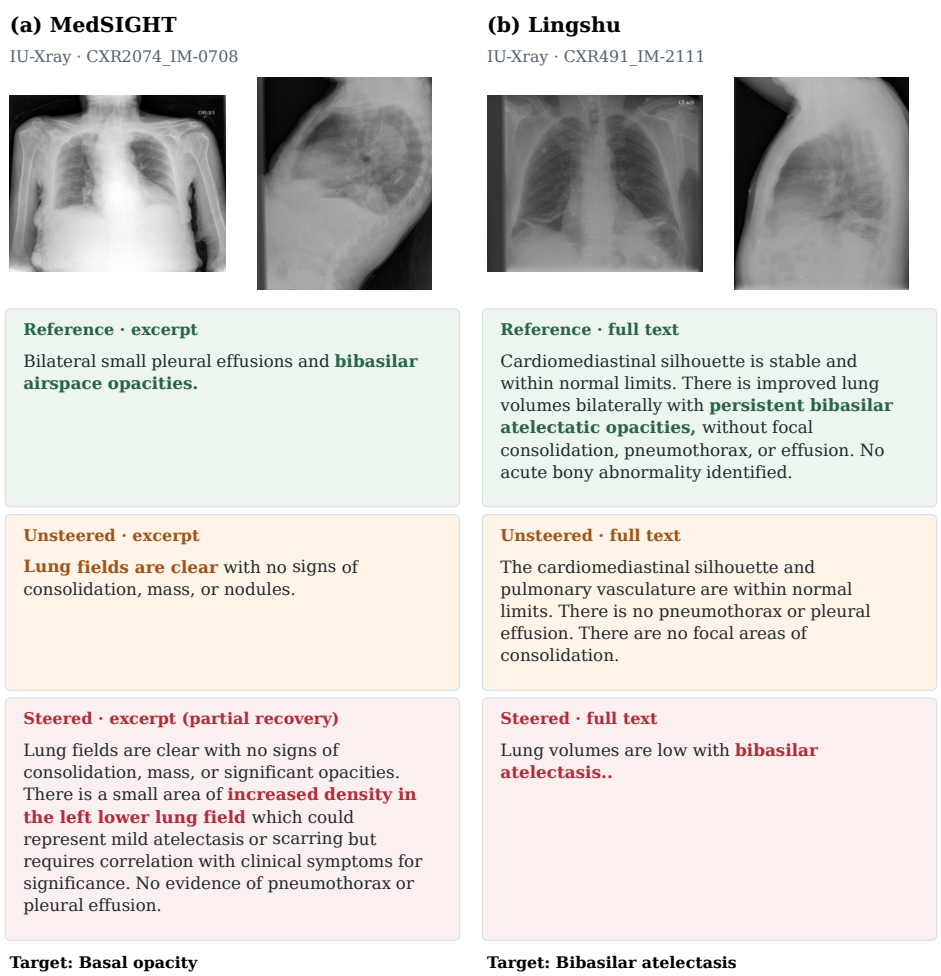


Figure 14: **Finding recovery in IU-Xray reports.** Green highlights the reference finding, amber marks the unsteered report, and red highlights the steered finding. **(a) MedSIGHT** changes from clear lungs to a left basal density, partially recovering the reference's bibasilar opacities. Its competing clear-lung statement and denial of the reference pleural effusions remain visible; the cause of the density is uncertain. **(b) Lingshu** adds the omitted bibasilar atelectasis. Its low-volume statement is not directly established by the reference. Text agreement is not independent image adjudication.

in English." Validation selects strength 1 with positive-gate-only application at the final prompt position and the next seven generation forwards. Decoding is greedy with a 128-token cap.

MedSIGHT uses a structured Findings/Impression prompt, strength 0.125, and negative-gate-only application at the final prompt position. Generation settings were selected on a balanced 20-study validation subset with a 128-token cap. To accommodate longer structured reports, the frozen setting was evaluated on the shared 100 studies at 256 tokens without further validation or reselection. Each model's baseline uses its corresponding prompt, token cap, and preprocessing.

E.4 Paired IU-Xray report examples

To illustrate how steering changes a generated report, Figure 14 presents two retrospectively selected finding recoveries from the shared 100-study IU-Xray cohort, using each model's setting in Table 4. MedSIGHT uses the structured-prompt, 256-token follow-up described in Appendix E.3; Lingshu uses the original report setting. Both input views are shown. MedSIGHT text consists of contiguous verbatim excerpts; Lingshu reports are reproduced in full. Neither reference contains masked spans. Complete outputs and the earlier failure cases remain in the result ledger. These examples illustrate successful finding-level changes using our steering directions.