CROSS-MODAL LATENT INFLUENCE: PATCH-LEVEL TRAINING SAMPLE ATTRIBUTION FOR VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Explaining a vision–language model (VLM) prediction requires tracing training-example influence to the multimodal evidence that mediates it, yet existing attribution methods largely operate only at the sample level. Recent latent-mediated influence methods provide feature-level attribution in language models, but lack the multimodal and spatial structure needed to ground influence in VLMs. We introduce `MOSAIC` (Multimodal Orthogonalized Sparse Autoencoders for Influence Characterization). It learns a shared sparse image–text latent space whose concepts are 1) sparse within each modality, 2) supported on contiguous image regions, 3) aligned to the phrases that describe them, and 4) approximately decorrelated. It then applies EK-FAC to decompose training-example influence through these concepts and localize their contributions to specific text phrases and image regions. We evaluate `MOSAIC` on Qwen and Gemma across captioning, chest X-ray understanding, and visual question answering. The learned features are more localized and coherent without degrading task behavior, and ablating influence-ranked latents changes model outputs more than ablating latents ranked by activation or frequency. We find that visual evidence can substantially reorganize training-data attribution while preserving more of the shared latent structure. `MOSAIC` thus connects *where model behavior comes from* with *what multimodal evidence carries that influence*.

1 INTRODUCTION

Understanding a visual prediction requires more than identifying salient pixels or patches: it requires linking the visual evidence a model relies on to the training examples that shaped that reliance. Vision attribution methods can localize prediction-relevant regions (Selvaraju et al., 2017; Chefer et al., 2021) (Fig. 1a), while data-attribution methods trace predictions to influential training examples (Koh & Liang, 2017; Park et al., 2023; Grosse et al., 2023). Yet, these two views are rarely connected, especially for vision–language models (VLMs). Even when data attribution successfully identifies an influential training example, it treats the image as a monolithic unit, answering only *which* example mattered, not *how* it mattered. To truly audit a VLM, we must drill down from the entire training sample to the specific concepts within it that actually drove the prediction.

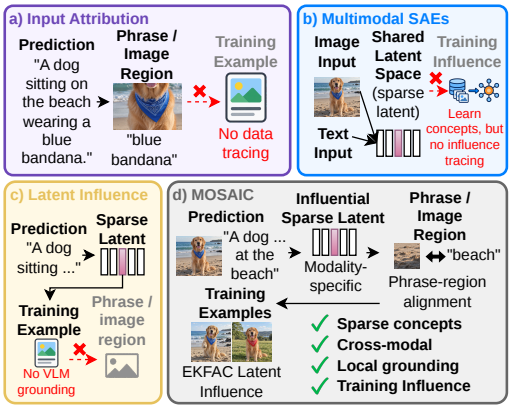


Figure 1: (a) Input attribution localizes evidence but lacks training-data specific attribution. (b) Multimodal SAEs expose internal concepts without tracing their training origins. (c) Latent influence traces training examples through features but lacks multimodal grounding. (d) `MOSAIC` connects influential training examples to sparse multimodal concepts and localized evidence.

VLMs make this finer-grained, concept-level attribution uniquely tractable. Unlike language-only models, influential features can be inspected along both *semantic* and *spatial* dimensions: a feature associated with a phrase can be directly checked against the image region on which it

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.

activates. This dual-sided inspection reveals not just the influential example, but *what concept carries its influence and where that concept appears*. This yields the multimodal attribution chain: training example ↔ multimodal concept ↔ phrase / image region.

Recent work on latent-mediated influence provides an important step toward such explanations in language models (Yu et al., 2026). It decomposes training-example influence through sparse latent features and maps those features back to tokens (Fig. 1c). Extending this formulation to VLMs, however, is not straightforward. Since VLM representations jointly encode image and text tokens, visual evidence carries explicit spatial structure rather than forming a linear sequence; further, the two modalities can have highly imbalanced token counts while still sharing latent directions. Existing multimodal SAEs, in contrast, can provide interpretable image–text features (Pach et al., 2025; Shen et al., 2025; Shu et al., 2026; Liao et al., 2026), but are not tailored for training-data attribution and auditing (Fig. 1b).

We introduce `MOSAIC` (Multimodal Orthogonalized Sparse Autoencoders for Influence Characterization), a framework for grounded latent-mediated influence in VLMs (Fig. 1d) by learning a latent representation whose features are both spatially grounded and cross-modally aligned, then attributing training-example influence through those features. `MOSAIC` first learns a shared TopK dictionary over image and text representations while preserving their distinct structure. To ensure that visual features correspond to coherent regions rather than scattered patches, it groups patches into regions, then imposes group and exclusive sparsity. To ensure that features are comparable across modalities, it aligns pooled phrase and region responses in the shared latent space and decorrelates latent activity, yielding a better-separated basis for attribution. Second, given this representation, an EK-FAC-based (Grosse et al., 2023) influence estimator decomposes each training example's effect over individual sparse features and projects those contributions back to the corresponding text phrases and image regions: prediction ↔ training example ↔ sparse latent ↔ phrase / image region.

`MOSAIC` (as illustrated in Fig. 2) enables questions that language-only attribution cannot pose. Do influential features correspond to localized cross-modal evidence? Are they causally important to the prediction? And how does visual evidence change the training examples through which a prediction is supported? We study these questions across two VLM architectures (Qwen and Gemma) and three tasks: image captioning, chest X-ray understanding, and visual question answering.

Our experiments show that 1) the structured latents are spatially localized and semantically grounded, 2) influence ranking identifies behaviorally consequential features, and 3) visual evidence substantially reshapes training-data attribution. First, the structured latent space yields substantially more localized multimodal features while largely preserving task performance: on GQA with Qwen, visual footprint (the fraction of image patches on which a feature activates) drops from 0.688 to 0.120, indicating that visual features concentrate on substantially smaller regions, with similar gains across tasks/backbones.

Second, influence-ranked latents are markedly more behaviorally consequential than activation-based alternatives: on GQA with Gemma, removing the top 256 influence-ranked latents flips 38.3% of predictions versus 8.3% under activation ranking, with the same trend across datasets and models. Together, these results support the learned latent decomposition as a meaningful basis for attribution.

Finally, visual evidence substantially reshapes training-data attribution. Removing the image drives training-example rank correlation as low as $\rho = -0.797$, while latent rankings remain positively correlated at up to $\rho = 0.795$. Thus, the influential examples can change dramatically even as substantially more of the underlying latent feature structure is preserved.

In summary, our contributions are threefold: (1) a **grounded training-data attribution framework for VLMs** that links training examples to localized phrases and image regions through a shared sparse latent basis; (2) an **attribution-oriented multimodal sparse autoencoder (SAE)** whose features are spatially coherent and cross-modally aligned; and (3) an **empirical characterization** showing that visual evidence can substantially reorganize training-data attribution while latent structure remains more stable and that the resulting features are behaviorally consequential and semantically grounded.

2 Related Work

Training-data attribution and latent influence. Training-data attribution methods, from influence functions to scalable approaches such as TRAK and approximations like EK-FAC (Koh & Liang,

2017; Park et al., 2023; Grosse et al., 2023), estimate how training examples affect model predictions. Recent work extends attribution toward more scalable or concept-level analyses, including sparse-recovery-based attribution, semantic concept influence, and concept-level attribution in generative models (Dagli et al., 2026; Kowal et al., 2026; Park et al., 2026). Most closely related, Yu et al. (2026) decomposes example influence through approximately orthogonal SAE features, exposing the latent features that mediate a prediction. We extend this view to VLMs, where influence must be resolved jointly over linguistic and spatially organized visual representations, enabling a hierarchy from influential training examples to multimodal concepts and localized evidence.

Visual and multimodal attribution. Visual attribution methods localize prediction-relevant evidence using gradients, attention, or relevance propagation (Selvaraju et al., 2017; Abnar & Zuidema, 2020; Chefer et al., 2021). Recent work further studies latent visual attribution, modality reliance, grounding, causal intervention, and mechanistic tracing in VLMs (Dreyer et al., 2025; Parcalabescu & Frank, 2025; Geigle et al., 2024; Li et al., 2025; Wang et al., 2026; Park et al., 2025). Together, these approaches identify or manipulate prediction-relevant visual evidence, modality contributions, or internal mechanisms, but do not connect them to the training examples that shaped the prediction. Since attention itself need not provide a faithful explanation (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019), we use frozen VLM attention only to construct phrase–region correspondences; attribution is instead determined by training-data influence through the learned latent representation.

Sparse multimodal representations. SAEs decompose dense activations into sparse, interpretable features (Cunningham et al., 2023; Gao et al., 2024). Recent multimodal extensions study shared image–text features, cross-modal alignment, hierarchical and compositional visual concepts, and spatially structured representations (Pach et al., 2025; Shen et al., 2025; Shu et al., 2026; Nasiri-Sarvi et al., 2026; Lou et al., 2025; Zaigrajew et al., 2025; Zhao et al., 2026; Lanza et al., 2026; Liao et al., 2026). However, these representations are not designed to mediate example-level training-data influence. In contrast, we learn a shared, position-resolved image–text latent basis with spatial coherence and phrase–region alignment, and decompose training-example influence through this basis to connect predictions, training examples, multimodal concepts, and localized visual and linguistic evidence.

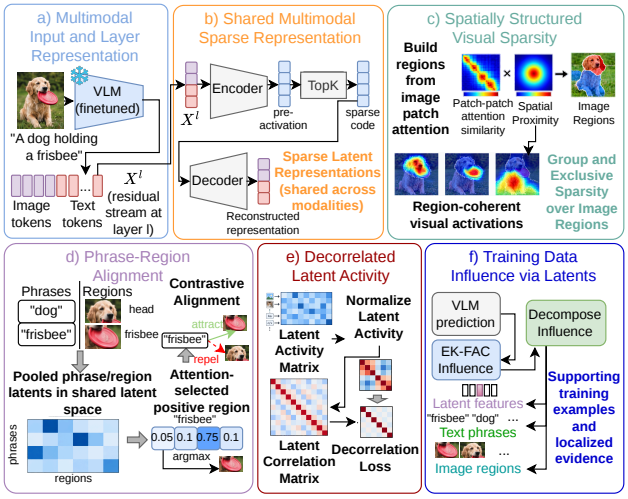


Figure 2: Overview of MOSAIC. (a) Intermediate VLM image and text representations. (b) Shared TopK SAE with a common sparse dictionary. (c) Spatially structured sparsity for region-coherent visual features. (d) Contrastive phrase–region alignment using frozen VLM attention. (e) Latent decorrelation for a better-separated attribution space. (f) EK-FAC influence tracing predictions to training examples, latents, phrases, and image regions.

3 METHOD

This section describes MOSAIC (Fig. 2). We first review the latent-mediated influence formulation on which our method builds and define the shared SAE representation (§3.1). We then introduce spatial constraints on visual features (§3.2) and phrase region alignment (§3.3). We aggregate latent attribution scores over image regions and text phrases (§3.4), using EK-FAC for scalable inverse curvature computation (§3.5). The SAE is optimized in two stages: learning a reconstructive dictionary, followed by spatial and cross-modal shaping. Full SAE configuration, layer placement, and optimization details are provided in Appendix H.

3.1 BACKGROUND AND MULTIMODAL LATENT BASIS

Following Yu et al. (2026), we split the VLM at layer l as $h_\theta = h_{\theta_2} \circ h_{\theta_1}$ and insert an SAE there. For example i, let $R_i \in \mathbb{R}^{T_i \times h}$ denote its sparse SAE activations, where r_{itj} is the activation of latent

j at sequence position t. The SAE decoder reconstructs the intermediate representation from R_i, which is then passed to the upper model h_{θ_2}.

For training example z_i and target example $z_\star$, define their gradients through the SAE-mediated forward path as $g_i(R_i) = \nabla_{\theta_2}\ell(z_i;\theta_2,R_i), \quad g_\star = \nabla_{\theta_2}\ell(z_\star;\theta_2,R_\star)$. Given curvature approximation C_{θ_2}, the estimated effect of upweighting z_i on the target loss is $I(z_i \to z_\star) = -g_\star^\top C_{\theta_2}^{-1} g_i(R_i)$.

To decompose this example-level influence over latent features, define $P_i(R_i) = g_\star^\top C_{\theta_2}^{-1} g_i(R_i)$. The contribution of latent j at position t is $A_{itj} = -r_{itj}\frac{\partial P_i(R_i)}{\partial r_{itj}}, \quad A_{ij} = \sum_{t=1}^{T_i} A_{itj}$. Here, the target gradient and curvature estimate are fixed while the training example's latent representation is varied, identifying which active features carry its influence on $z_\star$. This activation-weighted derivative gives a first-order attribution relative to an inactive-feature baseline, so inactive latents receive zero attribution. To make this decomposition multimodal and spatially localizable, we construct a latent basis over visual and linguistic representations. As illustrated in Fig. 2a–b, at layer l the VLM represents example i as a sequence of image- and text-token hidden states, $X_i^l = [x_{i1}^l \cdots x_{iT_i}^l]^\top \in \mathbb{R}^{T_i \times d}$, where $x_{it}^l \in \mathbb{R}^d$ is the hidden representation of one image or text token, T_i is the sequence length, and d is the VLM hidden dimension. We denote the image and text positions with $\mathcal{P}_{\text{img}}$ and $\mathcal{P}_{\text{txt}}$.

We apply the same TopK SAE position-wise to both modalities: $a_{it} = W_{\text{enc}}(x_{it}^l - b_{\text{pre}}) + b_{\text{enc}}, \quad r_{it} = \text{ReLU}(\text{TopK}_k(a_{it})), \quad \hat{x}_{it}^l = W_{\text{dec}} r_{it} + b_{\text{pre}}$. Here, $a_{it}, r_{it} \in \mathbb{R}^h$ are the latent pre-activation and sparse code, respectively, and $\hat{x}_{it}^l \in \mathbb{R}^d$ reconstructs the original hidden state. TopK_k retains the k largest entries before ReLU, yielding at most k active features per position. Stacking r_{it} across positions gives R_i above.

The shared encoder and decoder give image and text positions a common latent dictionary, while visual activations are further shaped by the spatial constraints in §3.2. Finally, because redundant latent coordinates make feature-level influence difficult to interpret, we encourage approximately decorrelated latent activity. This regularizes latent activation patterns rather than enforcing orthogonality of the overcomplete decoder dictionary. The exact decorrelation objective and SAE training details are provided in Appendix G.1.

3.2 Spatially Structured Visual Sparsity

Token-level sparsity alone does not ensure that a visual feature corresponds to a coherent image region. We therefore adapt the structured visual sparsity construction of Liao et al. (2026), grouping image patches using patch–patch attention, spatial proximity, and connectivity-constrained clustering.

For example i, let $\mathcal{G}_i = \{G_{i1}, \ldots, G_{iM_i}\}$ denote the partition of image-token positions produced by this clustering procedure, where $G_{im} \subseteq \mathcal{P}_{\text{img}}$ is region m and $M_i = |\mathcal{G}_i|$ is the resulting number of regions. Let $q_{itj} \in \{0,1\}$ denote whether latent j is selected by TopK at position t. We keep the TopK selection discrete in the forward pass and use an identity gradient approximation during backpropagation. Unlike Liao et al. 2026, we normalized regional support by the region size: $e_{ijm} = \left(\frac{1}{|G_{im}|}\sum_{t\in G_{im}} q_{itj}^2\right)^{1/2}$. Further, the spatial objective for an example with M_i regions is $\mathcal{L}_{\text{spatial}}^{(i)} = \lambda_{\text{group}} \sum_{j,m} e_{ijm} + \frac{\lambda_{\text{exclusive}}}{M_i}\sum_j \left(\sum_m e_{ijm}\right)^2$, where we additionally scale the exclusivity penalty term by M_i. These normalizations reduce sensitivity to region size and partition granularity, encouraging each latent's visual support to concentrate on a small number of coherent regions.

3.3 Phrase–Region Alignment

Spatial structure determines *where* a visual latent activates, but does not ensure that the same latent coordinate captures corresponding linguistic and visual evidence. We therefore introduce a phrase–region alignment objective that grounds shared SAE features across text phrases and image regions.

For example i, let Q_{iq} denote the text-token positions in phrase q, and let G_{im} denote the image-token positions in region m from §3.2. Given SAE pre-activations $[a_{i1} \cdots a_{iT_i}]^\top \in \mathbb{R}^{T_i \times h}$, we compute an RMS-pooled representation for any phrase or region g: $s_{ig} = \sqrt{\frac{1}{|g|}\sum_{t\in g} a_{it} \odot a_{it}}, \quad \bar{s}_{ig} =$

$s_{ig}/\|s_{ig}\|_2$, where the square root is applied element-wise and $s_{ig} \in \mathbb{R}^h$. We pool pre-activations before TopK and ReLU so that phrase–region alignment can use features even when they are not selected at every individual position.

To align phrases with image regions, we first define a latent-space similarity between phrase q and image region m, as $z_{iqm} = \bar{s}_{iQ_{iq}}^\top \bar{s}_{iG_{im}}/\tau$, where $\tau > 0$ is the temperature. To supervise this alignment, we must identify which region corresponds to each phrase. We use the frozen VLM's phrase-to-patch attention as a proxy for this ground truth. Let $A_i \in \mathbb{R}^{Q_i \times M_i}$ denote the matrix of VLM attention scores (affinities) between the Q_i phrases and M_i regions in example i. For each phrase q, we select the region with the highest attention score as its positive match: $\pi_i(q) = \arg\max_{m \in \{1,\dots,M_i\}} [A_i]_{qm}$. We then align each phrase with its matched region using a within-image contrastive objective: $\mathcal{L}_{\text{align}}^{(i)} = -\frac{1}{Q_i} \sum_{q=1}^{Q_i} \log \frac{\exp(z_{iq,\pi_i(q)})}{\sum_{m=1}^{M_i} \exp(z_{iqm})}$, where the remaining regions in the same image serve as negatives. Finally, the full cross-modal objective jointly minimizes alignment error, phrase-region magnitude mismatch, and excessive modality imbalance among strongly participating latents: $\mathcal{L}_{\text{cross}} = \lambda_{\text{al}}\mathcal{L}_{\text{align}} + \lambda_{\text{mag}}\mathcal{L}_{\text{mag}} + \lambda_{\text{bal}}\mathcal{L}_{\text{bal}}$. Definitions of the magnitude and balance terms are provided in Appendix G.1.1.

3.4 Multimodal Influence Decomposition

Latent influence identifies *which sparse features* carry the effect of a training example, but not where those features are expressed within that example. To recover this localization, in language models, we can compute the attribution of each individual token in the sequence via position-wise SAE activations (Yu et al., 2026). In a VLM, the sequence contains both text and image tokens, with each image token corresponding to a spatial location on the image patch grid. By calculating the position-wise attribution $A_{itj} = -r_{itj} \frac{\partial P_i(R_i)}{\partial r_{itj}}$, we preserve the exact mapping between the training example's influence and its manifestation at each specific word or image patch.

Text-token contributions can therefore be grouped into phrases, while image-token contributions can be visualized directly on their corresponding patches or aggregated over the spatial regions. For phrase $Q_{iq} \subseteq \mathcal{P}_{\text{txt}}$ and image region $G_{im} \subseteq \mathcal{P}_{\text{img}}$ of training example i, we aggregate $A_{ijq}^{\text{txt}} = \sum_{t \in Q_{iq}} A_{itj}$, $A_{ijm}^{\text{img}} = \sum_{t \in G_{im}} A_{itj}$. Together with $A_{ij} = \sum_{t=1}^{T_i} A_{itj}$, this organizes a training example's influence by latent feature and the localized linguistic and visual evidence through which it is expressed: prediction $\leftrightarrow$ training example $\leftrightarrow$ sparse latent $\leftrightarrow$ phrase / image patch or region. Details on the SAE-mediated forward path, sequence reduction, and patch-grid mapping are provided in Appendix G.

3.5 Scalable Influence Estimation with EK-FAC

The main computational bottleneck is the inverse-curvature product $C^{-1}g_i(R_i)$ in $I(z_i \to z_\star) = -g_\star^\top C^{-1} g_i(R_i)$. Constructing or inverting C directly is infeasible at VLM scale (we write $C = C_{\theta_2}$). We therefore use Eigenvalue-Corrected Kronecker-Factored Approximate Curvature (EK-FAC) (George et al., 2018; Grosse et al., 2023), which approximates the inverse curvature layer by layer while retaining structured second-order information. For a linear layer with input activation u and output gradient δ, K-FAC approximates its curvature block as $F \approx A \otimes S, \quad A = \mathbb{E}[uu^\top], \quad S = \mathbb{E}[\delta\delta^\top]$. Writing $A = Q_A \Lambda_A Q_A^\top$ and $S = Q_S \Lambda_S Q_S^\top$, EK-FAC retains the resulting Kronecker eigenbasis but replaces the factored eigenvalues with corrected values Λ_{EK} estimated from per-example gradients. For a layer gradient reshaped as V, the damped inverse-curvature product is approximated by $C^{-1}\,\text{vec}(V) \approx \text{vec}\left[Q_S \left(\frac{Q_S^\top V Q_A}{\Lambda_{\text{EK}} + \lambda}\right) Q_A^\top\right]$, where division is element-wise and $\lambda > 0$ is the damping coefficient.

Thus, rather than constructing the full curvature matrix, EK-FAC rotates each layer gradient into a structured eigenbasis, rescales each direction by its damped inverse curvature, and rotates it back. We apply this operation layerwise to compute the full curvature-corrected influence for each retrieved training example and, through $P_i(R_i)$, its latent decomposition from §3.4. The curvature statistics are estimated once and reused across training–test pairs. Training examples are ranked by the magnitude of the signed influence estimate, $|I(z_i \to z_\star)|$. Numerical comparisons with alternative curvature estimators and implementation details are provided in Appendices B, E, and G.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models. We evaluate Qwen3-VL-2B-Instruct (Team, 2025) and Gemma-4-E2B-it (Team, 2026), two open-weight VLMs from distinct model families. For each task, we finetune the backbone with LoRA, freeze the VLM, and train the SAE on its task-adapted intermediate representations. We place the SAE at a mid-to-late transformer layer to preserve downstream behavior while exposing semantically meaningful multimodal features. Additional details are provided in Appendix H and F.

Table 1: Task performance before and after SAE insertion. CIDEr for caption similarity to human references; accuracy is used for chest X-ray classification and VQA.

Dataset	Model	Metric	Base	FT	+SAE
VizWiz	Qwen	CIDEr	0.0738	0.7569	0.7471
VizWiz	Gemma	CIDEr	0.0686	0.5799	0.5793
CheXpert	Qwen	Acc.	0.3320	0.4290	0.4295
CheXpert	Gemma	Acc.	0.0040	0.4530	0.4560
GQA	Qwen	Acc.	0.5943	0.6513	0.6361
GQA	Gemma	Acc.	0.4019	0.5914	0.5867

Note: Base results for VizWiz and GQA use task-specific prompting. Without prompting: VizWiz CIDEr is 0.0001 for Qwen and 0.0003 for Gemma, while GQA accuracy is 0.0001 for Qwen and 0.0000 for Gemma.

Datasets and tasks. We evaluate three complementary vision–language settings. **VizWiz-Captions** tests open-ended captioning under challenging real-world imagery (Gurari et al., 2020). **CheXpert** tests three-way classification of radiographic findings as *absent*, *uncertain*, or *present* (Irvin et al., 2019). **GQA** evaluates compositional visual question answering over objects, attributes, and relations (Hudson & Manning, 2019).

4.2 GROUNDED MULTIMODAL REPRESENTATION

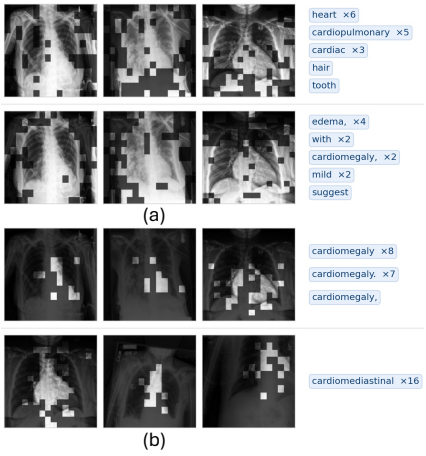


Figure 3: Comparison on CheXpert. Highlighted regions and words indicate latent activations. (a) Base SAE yields diffuse support and weaker radiological alignment. (b) Ours improves localization and clinically relevant alignment.

Before using the learned latents for attribution, we verify that the SAE preserves task behavior while providing the spatial and cross-modal structure required for grounded interpretation.

Behavior preservation. Table 1 compares each task-adapted VLM before and after SAE insertion. Across all datasets and backbones, performance changes only marginally, indicating that the sparse bottleneck preserves the behavior being explained. Full metrics in Appendix C.

Evaluation framework. We evaluate the learned basis along four dimensions: dictionary health, image–text balance, spatial localization, and semantic coherence. Visual footprint and contiguity measure how concentrated and spatially coherent a latent's visual support is, while MS Lift and causal grounding assess whether that support is semantically aligned with the corresponding linguistic representation. Full metric definitions and additional diagnostics are provided in Appendix D.

Objective ablation. Table 2 shows that structured sparsity and phrase–region alignment contribute complementary benefits. Structured sparsity reduces visual footprint from 0.312 to 0.120, while phrase–region alignment improves MS Lift from 0.0139 to 0.0191. Combining them yields the strongest overall representation, with footprint 0.083, contiguity 0.548, balance 26.7%, MS Lift 0.0214, and grounding 0.084. This suggests that spatial regularization primarily controls *where* a feature activates, while phrase–region alignment improves *what* that localized support represents across modalities. Additional comparisons can be found in Appendix K.

Generalization across tasks and backbones. The same trends hold across datasets and model families (Table 3). Visual footprint drops from 0.688 to 0.120 on GQA with Qwen and from 0.665 to 0.143 on VizWiz with Gemma, alongside improvements in dictionary health and modality balance. Fig. 3 shows the same effect qualitatively on CheXpert, where our latents exhibit sharper visual support and more relevant radiological terminology. Additional qualitative comparisons are provided in Appendix J.

4.3 VALIDATING LATENT INFLUENCE: NECESSITY AND SUFFICIENCY

Table 2: Latent-quality ablation on VizWiz with Qwen. Lower footprint and contiguity indicate more localized visual support; higher balance, MS Lift, and grounding indicate stronger multimodal and semantic structure. Grounding significance is assessed by permutation test.

Method	Dead↓	Bal.↑	Foot.↓	Contig.↓	MS↑	Ground.↑	p
Base SAE	74	22.5	0.312	0.641	0.0139	0.048	9.3×10^{-4}
+ Struct.	49	5.5	0.120	0.574	0.0049	0.012	3.1×10^{-4}
+ Align.	40	15.7	0.359	0.611	0.0191	0.039	1.2×10^{-3}
Full	**36**	**26.7**	**0.083**	**0.548**	**0.0214**	**0.084**	$\mathbf{2.9\times10^{-4}}$

Table 3: Latent quality across tasks and backbones. Parentheses report the fraction of active latents satisfying the 25% footprint threshold used for MS Lift. Visual MS is omitted for CheXpert.

Dataset	Model	Method	Dead↓	Bal.↑	Contig.↓	Foot.↓	MS Lift↑
VizWiz	Gemma	Base SAE	3642	10.1	0.302	0.665	0.0130 (0.05)
		Ours	**206**	**18.2**	**0.202**	**0.143**	**0.0233** (**0.96**)
CheXpert	Qwen	Base SAE	**23**	4.9	0.629	0.577	—
		Ours	55	**23.9**	**0.526**	**0.092**	—
	Gemma	Base SAE	315	7.3	0.432	0.956	—
		Ours	**158**	**38.7**	**0.340**	**0.178**	—
GQA	Qwen	Base SAE	24	12.0	0.603	0.688	0.0309 (0.20)
		Ours	**7**	**13.2**	**0.481**	**0.120**	**0.0394** (**1.00**)
	Gemma	Base SAE	5759	3.3	0.652	0.769	— (0.00)
		Ours	**184**	**9.4**	**0.261**	**0.246**	**0.0255** (**0.56**)

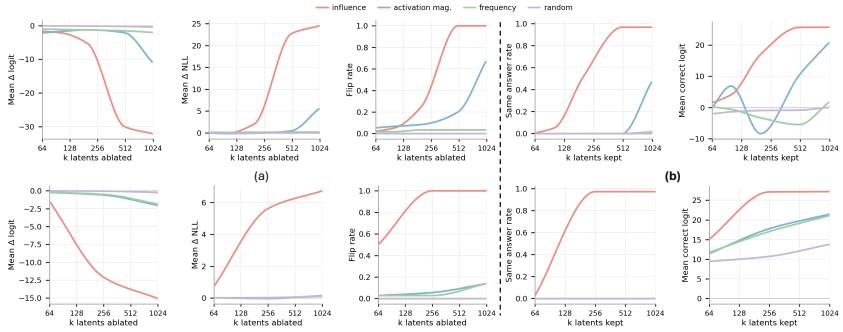


Figure 5: VizWiz necessity–sufficiency. Qwen [top], Gemma [bottom]; (a) necessity, (b) sufficiency.

Grounded features are useful for attribution only if they actually matter to the prediction. We therefore test whether high-influence latents are *necessary* for producing the output and *sufficient* to recover it.

For each test example, we aggregate influence across retrieved training examples and rank latents by this score. Necessity removes the top-k latents, while sufficiency keeps only the top-k. We compare against activation magnitude, firing frequency, and random rankings for $k \in [64, 1024]$. Necessity is evaluated by changes in correct-token logit, NLL, and prediction flips; sufficiency by same-answer rate and retained correct-token logit. Implementation details are in Appendix G.

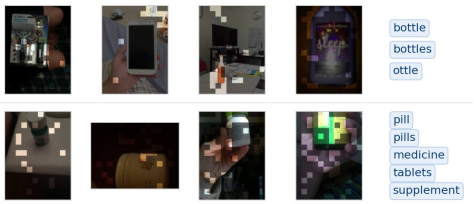


Figure 4: Latent breakdown for a VizWiz example with Qwen. The top two latents following Fig. 8 capture medicine- and bottle-related evidence; highlighted regions and words indicate representative activations.

Figs. 5–7 show consistent separation across datasets and backbones. Removing influence-ranked latents generally causes substantially larger prediction changes than removing activation-, frequency-, or randomly ranked features. For example, on GQA with Gemma at $k = 256$, influence-based removal flips 38.3% of predictions, compared with 8.3% under activation ranking. Similarly, the complementary sufficiency results show that the same features preserve substantially more of the original prediction when isolated. On GQA with Gemma, only 64 influence-ranked latents recover the original answer for 75.0% of examples, compared with 48.3% for activation magnitude. Similar trends hold on VizWiz and CheXpert. Thus, the latents surfaced by the influence estimator are not merely descriptive features: they carry important information for the model's prediction.

4.4 How Visual Evidence Shapes Training-Data Influence

The shared image–text latent space lets us ask whether removing visual evidence changes *which training examples* influence a prediction more strongly than it changes the *concepts* through which that influence is expressed. We evaluate this by recomputing influence for the same test examples with the image removed, while keeping the SAE and influence-estimation procedure fixed.

Across six model–dataset settings, absolute latent influence remains moderately to strongly correlated between the multimodal and text-only conditions, with Spearman ρ ranging from 0.543 to 0.795. In contrast, training-example attribution is substantially reorganized, with example-level correlations

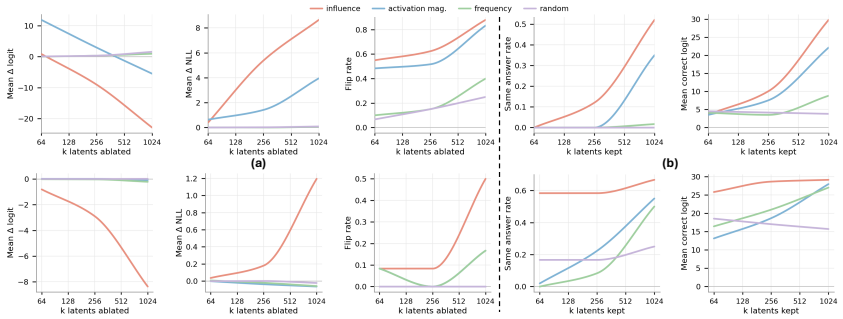


Figure 6: CheXpert necessity–sufficiency. Qwen [top], Gemma [bottom]; (a) necessity, (b) sufficiency.

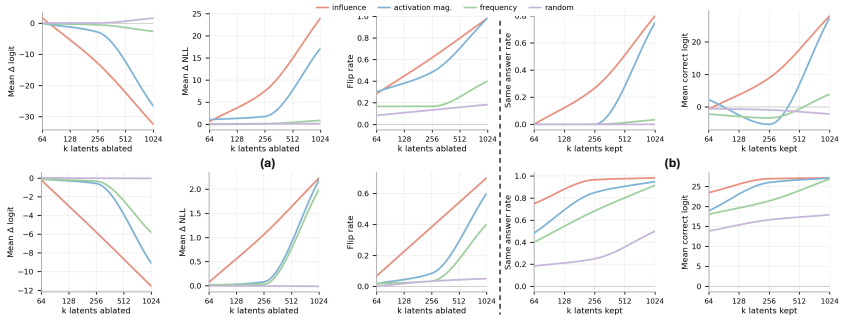


Figure 7: GQA necessity–sufficiency. Qwen [top], Gemma [bottom]; (a) necessity, (b) sufficiency.

ranging from -0.797 to -0.564. Figure 9 shows a representative example, where removing the image sharply reorders the influential training examples but maintains latent organization.

These results suggest that visual evidence primarily changes the *training-data support* underlying a prediction, while the latent concepts carrying influence remain comparatively more stable. The learned multimodal latent space provides a relatively stable conceptual interface, whereas the training examples associated with those concepts remain strongly conditioned on the available visual evidence. Additional analyses and results are provided in Appendix I.

4.5 Hierarchical Multimodal Attribution

The preceding experiments establish that the learned representation is grounded and that influence-ranked latents are behaviorally meaningful. We now examine what this decomposition reveals about individual predictions.

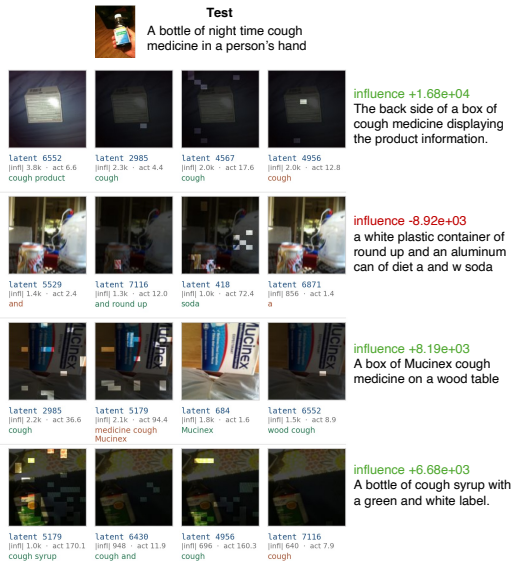


Figure 8: Training-example breakdown for a VizWiz example with Qwen. Training examples are ranked by influence; we show the top contributing latents, with peak image activations and caption words. Additional results are in the appendix.

Fig. 8 illustrates the full attribution hierarchy for a representative VizWiz example. Influential training examples contain prediction-relevant concepts such as cough medicine and related containers, while decomposing their influence through individual latents reveals finer visual and linguistic evidence, including medicine, bottle, and the wooden surface on which the object appears. The same latent contributions can be localized to both text phrases and image regions, exposing not only which training examples matter but also the grounded concepts through which their influence is expressed.

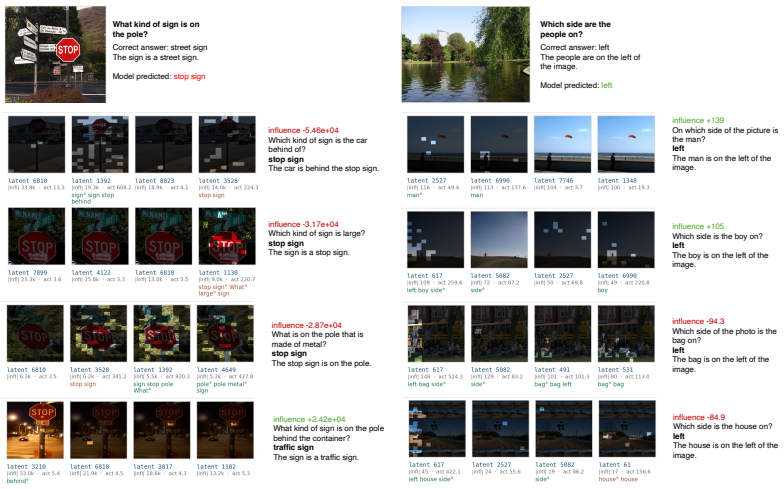


Figure 10: GQA influence decompositions. Incorrect prediction [left]. Correct prediction [right].

Fig. 4 provides the complementary test-conditioned view, directly ranking influential latents on the query representation. Here, the highest-ranked features capture medicine- and bottle-related evidence localized to the corresponding parts of the image. Together, the two views separate *where influence originates* from *what multimodal evidence carries it*. Additional test-latent decompositions and influential-training-example analyses are provided in Appendix J.

4.6 Attribution for Correct and Incorrect Predictions

We further examine whether the attribution hierarchy helps distinguish successful from failed predictions. Fig. 10 contrasts two representative GQA examples: one answered correctly by the model and one answered incorrectly. For each, we trace the prediction to its most influential training examples and decompose their influence through the highest-contributing multimodal latents.

In the correct case, the influential examples and latent concepts align with the relevant visual evidence and support the predicted answer. In the incorrect case, they instead concentrate on visually salient evidence associated with the model's mistaken prediction, revealing how the model can rely on the wrong grounded concept despite attending to relevant image content. This comparison shows how grounded influence can diagnose not only *what* supports a prediction, but also *why* that support leads to success or failure.

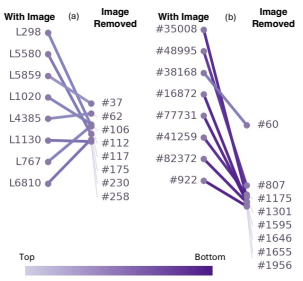


Figure 9: Effect of removing visual evidence on VizWiz influence rankings with Qwen. (a) Latent rankings remain relatively stable. (b) Training-example rankings change substantially.

5 Conclusion and Discussion

We introduced MOSAIC, a framework for tracing VLM predictions from influential training examples through sparse multimodal concepts to localized linguistic and visual evidence. By combining a shared image–text latent space with spatially coherent visual structure, phrase–region alignment, and latent-mediated influence, MOSAIC provides a grounded view of how training data shapes individual predictions. Across two VLM families and three tasks, the learned representation preserves model behavior while yielding more localized and semantically coherent features, and influence-ranked latents are substantially more behaviorally consequential than activation- or frequency-based alternatives. Our analysis further shows that visual evidence can strongly reorganize which training examples matter while preserving substantially more of the underlying latent structure. These results provide a path toward more fine-grained and auditable training-data attribution in multimodal models. Limitations and future work are discussed in Appendix A.

AI USE STATEMENT

We used generative AI tools to assist in the manuscript writing process. The authors have reviewed all uses of generative AI and content overall. All scientific work was done and verified by the authors. We take responsibility for the final content and claims.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details needed to reproduce our results throughout the main text and appendices. Section 3 specifies the architecture, training objectives, and latent-influence formulation, while Section 4.1 describes the evaluated models, datasets, and task setup. Appendix G provides additional details for latent-mediated influence computation and intervention-based evaluation, Appendix G.1 gives SAE configuration, layer placement, and optimization details. Metric definitions and additional evaluation diagnostics are provided in Appendix D.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.385.

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 387–396, 2021.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Rishit Dagli, Abir Harrasse, Luke Zhang, Florent Draye, Amirali Abdullah, Bernhard Schölkopf, and Zhijing Jin. Stride: Training data attribution via sparse recovery from subset perturbations, 2026. URL https://arxiv.org/abs/2606.05165.

Maximilian Dreyer, Lorenz Hufe, Jim Berend, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. From what to how: Attributing clip's latent components reveals unexpected semantic reliance, 2025. URL https://arxiv.org/abs/2505.20229.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024. URL https://arxiv.org/abs/2406.04093.

Gregor Geigle, Radu Timofte, and Goran Glavaš. Does object grounding really reduce hallucination of large vision-language models?, 2024. URL https://arxiv.org/abs/2406.14492.

Thomas George, César Laurent, Xavier Bouthillier, Nicolas Ballas, and Pascal Vincent. Fast approximate natural gradient descent in a kronecker-factored eigenbasis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 9573–9583, Red Hook, NY, USA, 2018. Curran Associates Inc.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilė Lukošiūtė, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. Studying large language model generalization with influence functions, 2023. URL https://arxiv.org/abs/2308.03296.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII*, pp. 417–434, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58519-8. doi: 10.1007/978-3-030-58520-4_25. URL https://doi.org/10.1007/978-3-030-58520-4_25.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering, 2019. URL https://arxiv.org/abs/1902.09506.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.3301590. URL https://doi.org/10.1609/aaai.v33i01.3301590.

Sarthak Jain and Byron C. Wallace. Attention is not explanation, 2019. URL https://arxiv.org/abs/1902.10186.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1885–1894. JMLR.org, 2017.

Matthew Kowal, Goncalo Paulo, Louis Jaburi, Tom Tseng, Lev E McKinney, Stefan Heimersheim, Aaron David Tucker, Adam Gleave, and Kellin Pelrine. Concept influence: Leveraging interpretability to improve performance and efficiency in training data attribution, 2026. URL https://arxiv.org/abs/2602.14869.

Sergio Lanza, Jae Hee Lee, and Stefan Wermter. Extraction and analysis of multimodal concepts in vision language models through sparse autoencoders, 2026. URL https://arxiv.org/abs/2606.21197.

Wei Li, Zhen Huang, Houqiang Li, Le Lu, Yang Lu, Xinmei Tian, Xu Shen, and Jieping Ye. Visual evidence prompting mitigates hallucinations in large vision-language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4048–4080, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.205. URL https://aclanthology.org/2025.acl-long.205/.

Weiduo Liao, Yunqiao Yang, and Ying Wei. When structured sparse autoencoders learn consistent concepts across modalities, 2026. URL https://arxiv.org/abs/2607.08605.

Hantao Lou, Changye Li, Jiaming Ji, and Yaodong Yang. Sae-v: Interpreting multimodal models for enhanced alignment, 2025. URL https://arxiv.org/abs/2502.17514.

Ali Nasiri-Sarvi, Hassan Rivaz, and Mahdi S. Hosseini. Sparc: Concept-aligned sparse autoencoders for cross-model and cross-modal interpretability, 2026. URL https://arxiv.org/abs/2507.06265.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models, 2025. URL https://arxiv.org/abs/2504.02821.

Letitia Parcalabescu and Anette Frank. Do vision language decoders use images and text equally? how self-consistent are their explanations?, 2025. URL https://arxiv.org/abs/2404.18624.

Sangha Park, Seungryong Yoo, Jisoo Mok, and Sungroh Yoon. Save: Sparse autoencoder-driven visual information enhancement for mitigating object hallucination, 2025. URL https://arxiv.org/abs/2512.07730.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Mądry. Trak: attributing model behavior at scale. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Yonghyun Park, Chieh-Hsin Lai, Satoshi Hayakawa, Yuhta Takida, Naoki Murata, Wei-Hsiang Liao, Woosung Choi, Kin Wai Cheuk, Junghyun Koo, and Yuki Mitsufuji. Concept-trak: Understanding how diffusion models learn concepts through concept-level attribution, 2026. URL https://arxiv.org/abs/2507.06547.

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Teodora Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7(1):119–130, Jan 2025. ISSN 2522-5839. doi: 10.1038/s42256-024-00965-w. URL https://doi.org/10.1038/s42256-024-00965-w.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Shufan Shen, Junshu Sun, Qingming Huang, and Shuhui Wang. Vl-sae: Interpreting and enhancing vision-language alignment with a unified concept set, 2025. URL https://arxiv.org/abs/2510.21323.

Huizhen Shu, Xuying Li, Hongxu Lin, Wenjie Sun, and Hui Li. Steering vision-language models with joint sparse autoencoders, 2026. URL https://arxiv.org/abs/2606.25657.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Qidong Wang, Junjie Hu, and Ming Jiang. V-seam: Visual semantic editing and attention modulating for causal interpretability of vision-language models, 2026. URL https://arxiv.org/abs/2509.14837.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *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. 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002/.

Shixing Yu, Promit Ghosal, and Kyra Gan. Correcting influence: Unboxing llm outputs with orthogonal latent spaces, 2026. URL https://arxiv.org/abs/2605.12809.

Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. Interpreting clip with hierarchical sparse autoencoders, 2025. URL https://arxiv.org/abs/2502.20578.

Yusong Zhao, Hengyi Wang, Tanuja Ganu, Akshay Nambi, and Hao Wang. Cascaded sparse autoencoders learn multi-level visual concepts in multimodal llms, 2026. URL https://arxiv.org/abs/2606.16193.

A Limitations and Future Work

Our approach depends on the quality of the learned SAE representation: latent features may remain partially entangled, and attribution can depend on the chosen layer and dictionary capacity. Phrase–region alignment also relies on pretrained attention as a correspondence signal, which may be noisy or incomplete. In addition, EK-FAC and latent-mediated influence remain approximations to the effect of full retraining. Future work could study multi-layer or hierarchical latent spaces, attention-independent grounding objectives, and stronger validation against retraining-based influence. Extending grounded latent influence to larger VLMs, video and other modalities, and applications such as data debugging, hallucination analysis, and model editing is also promising.

B Curvature Approximation Study

Computing the representation-mediated influence requires applying an inverse-curvature operator to the test gradient. Explicitly forming this matrix is infeasible even for the upper-model parameter subset used for influence computation, which contains approximately 6×10^8 parameters. We therefore evaluated several scalable curvature approximations before adopting EK-FAC. The alternatives exhibit qualitatively different failure modes, ranging from unstable linear solves to curvature estimates that are numerically well behaved but provide little meaningful geometric correction.

Raw Hessian. We first considered implicit Hessian–vector products combined with conjugate gradients (CG). This approach is unreliable in our setting because the empirical Hessian is strongly indefinite, whereas standard CG assumes a positive-definite system. The solver repeatedly encounters negative curvature and terminates prematurely.

This instability translates directly into unstable influence estimates. Across `cg_iters` $\in \{10, 40, 160\}$, the maximum absolute influence changes from

$$0.19 \;\rightarrow\; 0.35 \;\rightarrow\; 3.31,$$

and the solutions obtained with 10 and 160 iterations share no examples among their respective top-10 training examples. Thus, the resulting attribution depends strongly on the optimization trajectory rather than defining a stable inverse-Hessian estimate.

Generalized Gauss–Newton. The generalized Gauss–Newton (GGN) matrix avoids the negative-curvature issue because it is positive semidefinite by construction. However, its linear systems remain difficult to solve in our setting. Even under substantial damping, residuals fail to decrease reliably, while each solve is substantially more expensive than the alternatives considered here. Thus, although GGN provides a more well-behaved curvature object than the raw Hessian, it does not offer a practical inverse-curvature estimator at the scale required for our experiments.

Empirical Fisher. The empirical Fisher is attractive because it is positive semidefinite and admits an efficient low-rank inverse. For a curvature cohort of m examples,

$$F = \frac{1}{m}\sum_{i=1}^{m} g_i g_i^\top. \tag{1}$$

With damping, the resulting system is numerically stable and can be solved efficiently through its low-rank representation. Its main limitation in our setting, however, is not numerical but geometric.

In particular, the cohort gradients are nearly orthogonal: the mean pairwise cosine similarity is -0.002. As a result, each additional example contributes an almost independent direction rather than reinforcing a shared low-dimensional curvature structure. With only $m = 32$ curvature examples in a parameter space of approximately 6×10^8 dimensions, the empirical Fisher therefore covers an extremely small fraction of the directions relevant to the test gradient. Empirically, doubling the cohort size increases the captured gradient mass by only approximately $1.3\times$, indicating that this coverage improves slowly with additional curvature examples.

This low-rank geometry makes the damped inverse close to a scaled identity outside the cohort span. We quantify this using the alignment between the inverse-curvature product and the original test

Table 4: Full VizWiz captioning results. We report CIDEr, METEOR, BLEU-1–4, and ROUGE-L. "Prompted" uses the task-specific captioning prompt, while "zero-shot" denotes evaluation without task-specific prompting.

Model	Setting	CIDEr↑	METEOR↑	B-1↑	B-2↑	B-3↑	B-4↑	R-L↑
Qwen	Finetuned, no SAE	0.7569	0.2641	0.7076	0.5171	0.3736	0.2660	0.4767
Qwen	+ SAE	0.7471	0.2612	0.7099	0.5180	0.3737	0.2659	0.4758
Qwen	Base (prompted)	0.0738	0.2385	0.3452	0.2070	0.1205	0.0706	0.3079
Qwen	Base (zero-shot)	0.0001	0.0026	0.0001	0.0000	0.0000	0.0000	0.0108
Gemma	Finetuned, no SAE	0.5799	0.2405	0.6608	0.4558	0.3068	0.2030	0.4419
Gemma	+ SAE	0.5793	0.2389	0.6622	0.4557	0.3057	0.2016	0.4416
Gemma	Base (prompted)	0.0686	0.1747	0.2902	0.1564	0.0826	0.0449	0.2489
Gemma	Base (zero-shot)	0.0003	0.0156	0.0230	0.0010	0.0000	0.0000	0.0265

gradient,

$$\mathrm{CurvAlign} = \cos\big((F + \lambda I)^{-1} g_{\mathrm{test}}, g_{\mathrm{test}}\big)\,. \tag{2}$$

For the empirical Fisher, we obtain $\mathrm{CurvAlign} = 0.987$, showing that the curvature transformation leaves the test-gradient direction almost unchanged. In this regime, the resulting influence score approaches a rescaled first-order gradient-similarity score rather than substantially incorporating second-order geometry.

Notably, standard numerical checks do not expose this failure mode. The Woodbury solve attains residual ratios of approximately 10^{-6}, the associated Gram matrix has condition number 2.55, and influence rankings remain highly stable across disjoint curvature cohorts (Spearman $\rho = 0.993$). These diagnostics confirm that the low-rank system is solved accurately, but they do not establish that the curvature approximation is informative. Indeed, the high $\mathrm{CurvAlign}$ value shows that the empirical Fisher is stable largely because most directions remain dominated by the damping term.

Thus, in our setting, the empirical Fisher is not rejected because of poor conditioning or solver error. Rather, the near-orthogonality of per-example gradients makes a small-cohort empirical Fisher too low-rank to induce a meaningful transformation of the test gradient.

EK-FAC. We use EK-FAC for the final influence computation; the approximation itself is described in the main text. EK-FAC provides a structured layerwise curvature approximation while avoiding the severe rank restriction of cohort-based curvature estimates. This is particularly important in our setting, where per-example gradients are nearly orthogonal and a small collection of training gradients spans only a negligible portion of the full parameter space. By instead modeling curvature through layerwise activation and gradient statistics, EK-FAC captures curvature structure beyond the span of individual examples while remaining computationally tractable at model scale.

We additionally validate the EK-FAC approximation in controlled small-scale settings where the target curvature can be explicitly materialized. When constructed at the example level, consistent with our influence formulation, EK-FAC achieves cosine fidelity of 0.997–0.9996 across dense, ragged, and masked conditions. This high agreement indicates that the structured approximation closely preserves the curvature transformation relevant to our influence estimates while remaining tractable at model scale.

These results support EK-FAC as a practical approximation for our setting: it provides a scalable, non-degenerate curvature transformation while preserving the example-level geometry required by our influence formulation. We therefore use EK-FAC for all influence results reported in the main paper.

C Full VizWiz Captioning Results

Table 4 reports the complete set of captioning metrics for VizWiz. LoRA finetuning substantially improves performance over both prompted and unprompted base models, while SAE insertion results in only minor changes across metrics, indicating strong preservation of task-relevant information.

D Additional Representation Evaluation

We provide additional details for the latent-quality metrics used in Sec. 4.2, including dictionary health, modality balance, visual footprint, contiguity, MS Lift, and causal grounding. We additionally report diagnostics of latent activity correlation and layer placement to verify that the selected SAE interface provides a well-separated representation while preserving downstream behavior.

D.1 Latent Quality Metrics

Here we provide precise definitions that are useful for interpreting these quantities but are omitted from the main tables for compactness. All statistics are computed on a fixed held-out set of 1,500 training examples, independently of the influence pipeline.

Dead latents. A latent is considered dead if it never activates on any of the 1,500 evaluation examples.

Modality balance. For latent j and modality $m \in \{\text{img}, \text{txt}\}$, we define the token-normalized activation rate

$$r_j^m = \frac{\text{mass}_j^m}{N_{\text{tokens}}^m}, \tag{3}$$

and image share

$$s_j = \frac{r_j^{\text{img}}}{r_j^{\text{img}} + r_j^{\text{txt}}}. \tag{4}$$

A live latent is considered balanced when $s_j \in [0.35, 0.65]$. Token normalization prevents modality share from being dominated by differences in image and text sequence length.

Spatial footprint. Spatial footprint measures the fraction of an individual image's patch grid occupied by the latent.

Spatial contiguity. Raw connected-component counts depend strongly on support size, so we normalize each latent's component count by that of a size-matched random mask on the same patch grid,

$$C_{\text{rel}} = \frac{C(\mathcal{S})}{\mathbb{E}_{\widetilde{\mathcal{S}}:|\widetilde{\mathcal{S}}|=|\mathcal{S}|}[C(\widetilde{\mathcal{S}})]}. \tag{5}$$

Values below 1 indicate support that is more spatially contiguous than expected by chance. We treat contiguity as supplementary to footprint rather than as a standalone localization measure.

Visual monosemanticity. Following the MonoSemanticity Score (MS) of Pach et al. (2025), we assess whether the regions activating a latent are semantically coherent under an independent reference encoder. We adapt MS to localized VLM features by evaluating the image patches on which the latent itself fires using DINOv2.

Raw MS is strongly confounded by footprint size. We therefore compare the observed support $\mathcal{S}_j$ with a random support of identical size drawn from the same images and report

$$\text{MSLift}_j = \text{MS}(\mathcal{S}_j) - \mathbb{E}\left[\text{MS}(\widetilde{\mathcal{S}}_j)\right]. \tag{6}$$

The main text reports this null-corrected lift rather than raw MS. As the support approaches full-image coverage, the matched random support increasingly overlaps the observed support and the comparison becomes degenerate. We therefore score only latents whose median footprint covers less than 25% of the patch grid.

Why visual MS is omitted for CheXpert. For CheXpert, RAD-DINO (Pérez-García et al., 2025) produces a substantially higher size-matched similarity baseline than DINOv2 does on natural images. Because chest radiographs exhibit highly stereotyped global anatomy, arbitrary regions can already be strongly similar across examples. The assumption that a meaningful latent region should be more self-consistent than its surroundings is therefore poorly matched to this domain, and we omit visual MS from the CheXpert evaluation.

Table 5: Orthogonality of Qwen representations before and after SAE encoding at layer $\ell = 15$ ($d_{\text{latent}} = 8192$, $k = 128$).

Dataset	Metric	Text	Pre-latent	SAE latent
VizWiz	Near-orthogonal ($\lvert\rho\rvert < 0.1$) ↑	7.6%	20.8%	**99.5%**
	Mean $\lvert\text{offdiag}(G)\rvert$ ↓	0.798	0.280	**0.0020**
	Mean $\text{offdiag}(G)^2$ ↓	0.751	0.115	**0.0008**
	Stable rank (raw / norm.)	1.16/0.0053	4.01/0.0020	9.93/0.0012
CheXpert	Near-orthogonal ($\lvert\rho\rvert < 0.1$) ↑	24.9%	25.4%	**99.2%**
	Mean $\lvert\text{offdiag}(G)\rvert$ ↓	0.510	0.239	**0.0036**
	Mean $\text{offdiag}(G)^2$ ↓	0.424	0.086	**0.0010**
	Stable rank (raw / norm.)	1.54/0.0044	4.71/0.0023	17.25/0.0021

Causal grounding. For each latent j, we ablate it only at image patches where it is active and measure the change in negative log-likelihood of its top-associated word. We define

$$\text{share}(a) = \frac{\text{NLL}(a) - \text{NLL}_{\text{full}}}{\text{NLL}_{\text{no-image}} - \text{NLL}_{\text{full}}},$$

where $\text{NLL}_{\text{no-image}}$ is obtained by zeroing all SAE latents at all image positions. The causal grounding effect is

$$\text{GroundingEffect}_j = \text{share}(j) - \text{share}(c(j)),$$

where $c(j)$ is a control latent with similar active-patch count, ablated at its own active patches in the same image. We report the mean effect as `mean_solo_effect` together with `solo_effect_pvalue`. We exclude examples with image dependence below 0.05 and latents active on fewer than two image patches.

D.2 Latent-Space Orthogonality

Latent-mediated influence is easiest to interpret when individual SAE coordinates are weakly correlated, so that attribution to one latent is not simply a rotated description of a highly redundant subspace. This motivation is closely related to the approximately orthogonal latent basis used for influence attribution by Yu et al. (2026). We therefore evaluate whether our learned sparse representation becomes more orthogonal than the dense VLM activations from which it is constructed.

For Qwen, we measure orthogonality at three representation levels: the dense activations entering the SAE (*pre-latent*), the k-sparse SAE code (*SAE latent*), and the input text embeddings (*text*) as a reference. For each representation, we center and normalize the features, form the corresponding Gram matrix, and report the mean absolute and mean squared off-diagonal entries, the fraction of near-orthogonal pairs satisfying $\lvert\rho\rvert < 0.1$, and the stable rank

$$\text{srank}(A) = \frac{\|A\|_F^2}{\|A\|_2^2}. \tag{7}$$

Table 5 reports results for our full SAE at layer $\ell = 15$.

The SAE substantially decorrelates the representation on both datasets. On VizWiz, the fraction of near-orthogonal pairs increases from 20.8% in the dense representation to 99.5% after SAE encoding, while the mean absolute off-diagonal correlation decreases from 0.280 to 0.0020. CheXpert shows the same pattern, with near-orthogonal pairs increasing from 25.4% to 99.2% and mean absolute correlation decreasing from 0.239 to 0.0036. Thus, despite the different visual domains, the learned sparse codes consistently provide a substantially less correlated coordinate system for subsequent per-latent influence analysis.

Interpretation. The near-orthogonal fraction and off-diagonal Gram statistics provide the most direct comparisons because they measure correlation between coordinates and are less sensitive to representation width. The raw stable rank should be interpreted more cautiously: the SAE latent space is 8192-dimensional, compared with 2048 dimensions for the pre-latent representation, so its absolute

Table 6: Sensitivity of Gemma on GQA to SAE placement.

Setting	SAE Layer	GQA Acc.
Base model	–	0.4019
Fine-tuned model	–	0.5914
Fine-tuned + SAE	19	0.5867
Fine-tuned + SAE	12	0.3315

value is not directly comparable across spaces. Indeed, after normalization by the maximum attainable rank, the ordering can change. We therefore use stable rank only as a supplementary diagnostic and base our orthogonality claim on the near-orthogonal fraction and off-diagonal correlations.

The text representation is likewise included only as a reference. Its Gram statistics are computed over the token axis rather than the SAE feature axis and therefore should not be interpreted as a direct intermediate point between the dense and sparse representations. Our main conclusion is restricted to the comparable pre-latent and SAE-latent representations: SAE encoding substantially reduces coordinate correlation, providing a near-orthogonal basis better suited to latent-mediated influence attribution.

D.3 Sensitivity to SAE Layer Placement

We evaluate SAE placement sensitivity on Gemma for GQA using the same 8192-dimensional SAE with $k = 128$. While the primary layer-19 placement preserves fine-tuned performance, moving the SAE to layer 12 reduces accuracy to 0.3315.

This approximately 26-point drop shows that SAE placement materially affects representation fidelity as shown in Table 6, dropping below base performance. We therefore use layer 19 for the main Gemma experiments, where the SAE preserves downstream behavior while providing the latent basis used for attribution.

E Gradient-Based Candidate Retrieval

Computing full latent-mediated influence for every training–test pair is costly. Following Yu et al. (2026), we first retrieve candidate training examples using gradient similarity and apply the full curvature-aware influence estimator only to the retained set. For test example z_{test} and training example z_i, we compute $s_i = \langle \nabla_\theta \mathcal{L}(z_i;\theta), \nabla_\theta \mathcal{L}(z_{\text{test}};\theta) \rangle$ and retain $\mathcal{C}_K(z_{\text{test}}) = \text{TopK}_i\ s_i$. Gradient similarity is used only for candidate retrieval; all reported training-example rankings are determined by the latent-mediated EK-FAC influence estimator.

F Computational Scope and Model Scale

All SAE training and influence analyses are run on a single NVIDIA RTX PRO Blackwell GPU. At this setup, the approximately 2B-parameter models used in our experiments are already near the practical memory and compute limit for the full pipeline, which requires per-example gradients, EK-FAC statistics, and latent-resolved influence computation. We therefore restrict evaluation to this scale and leave larger VLMs to future work.

G Latent-Mediated Influence Implementation

SAE-mediated forward path. The SAE is inserted directly into the VLM forward computation at the selected intermediate layer. The upper model receives the reconstructed representation

$$\hat{X}_i^l = W_{\text{dec}} R_i + b_{\text{pre}},$$

rather than the original residual stream. Thus, all influence computations, latent interventions, and downstream evaluations are performed through the same SAE-mediated representation used for attribution.

Sequence reduction. For multimodal example i, the SAE representation is sequence-resolved, $R_i \in \mathbb{R}^{T_i \times h}$. For the global training-example influence computation, we reduce derivatives over valid sequence positions to obtain a common h-dimensional latent representation. Token-resolved activations are retained for the subsequent phrase- and region-level localization in Sec. 3.4. This reduction provides a consistent latent representation across variable-length image–text sequences for curvature and example-level influence, while the original sequence-resolved activations and derivatives are retained for phrase-, patch-, and region-level localization.

Position-resolved multimodal localization. Although example-level influence is computed using EK-FAC in model-parameter space, we decompose this influence through the SAE latent representation. For the global latent decomposition, derivatives are aggregated over valid sequence positions to obtain an h-dimensional latent representation, while the sequence-resolved SAE activations $R_i \in \mathbb{R}^{T_i \times h}$ and their position-wise derivatives are retained for localization. Thus, for latent j, the contribution

$$A_{itj} = -r_{itj} \frac{\partial P_i}{\partial r_{itj}}$$

remains indexed by the original sequence position t.

For text positions, t maps directly to the corresponding token and can be aggregated over phrase spans Q_q. For image positions, we retain the VLM's mapping from each visual token to its location on the image patch grid. We therefore reshape $\{A_{itj} : t \in \mathcal{P}_{\text{img}}\}$ back onto that grid to obtain a patch-level attribution map for latent j. Region-level attribution is obtained by summing the patch contributions belonging to region G_m:

$$A^{\text{img}}_{ijm} = \sum_{t \in G_m} A_{itj}, \qquad A^{\text{txt}}_{ijq} = \sum_{t \in Q_q} A_{itj}.$$

The patch-level maps are used for visualization, while phrase- and region-level aggregates provide the grouped multimodal decomposition reported in the main text.

Candidate retrieval. We use the gradient-based filtering procedure described in Appendix E before evaluating full EK-FAC influence.

G.1 SAE and Decorrelation Details

For completeness, we summarize the full representation-learning objective used to train the multimodal SAE. For a minibatch $\mathcal{B}$, the objective is

$$\mathcal{L}_{\text{SAE}} = \mathcal{L}_{\text{rec}} + \lambda_{\perp}\mathcal{L}_{\perp} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}} + \lambda_{\text{lb}}\mathcal{L}_{\text{lb}} + \lambda_{\text{group}}\mathcal{L}_{\text{group}} + \lambda_{\text{exclusive}}\mathcal{L}_{\text{exclusive}} + \mathcal{L}_{\text{cross}}.$$

The load-balancing term $\mathcal{L}_{\text{lb}}$ discourages activity from collapsing onto a small subset of the dictionary.

Reconstruction objective. The SAE is trained with

$$\mathcal{L}_{\text{rec}} = \frac{1}{T} \sum_{t=1}^{T} \|x^l_t - \hat{x}^l_t\|_2^2.$$

The VLM remains frozen throughout SAE training.

Decorrelated latent activity. For latent j, let $r_j = [r_{1j} \ \cdots \ r_{Tj}]^\top$ and define its centered, normalized activity as

$$\tilde{r}_j = \frac{r_j - \bar{r}_j \mathbf{1}}{\|r_j - \bar{r}_j \mathbf{1}\|_2 + \epsilon}.$$

We penalize off-diagonal correlations,

$$\mathcal{L}_{\perp} = \frac{1}{h(h-1)} \sum_{i \neq j} (\tilde{r}_i^\top \tilde{r}_j)^2.$$

This encourages approximately decorrelated latent activity without imposing orthogonality on the decoder dictionary.

G.1.1 Additional Phrase–Region Alignment Details

Section 3.3 introduces the phrase–region alignment objective used to ground shared SAE features across image and text. Here we provide the additional magnitude- and modality-balance terms used in the full cross-modal objective.

Response-magnitude matching. Contrastive alignment constrains the direction of normalized phrase and region representations but does not control their relative activation scale. For example b, we define the ratio between average region and phrase response magnitudes as $\rho_b = \frac{\frac{1}{|\mathcal{R}_b|}\sum_{r\in\mathcal{R}_b}\|s_r\|_2}{\frac{1}{|\mathcal{P}_b|}\sum_{p\in\mathcal{P}_b}\|s_p\|_2}$. We penalize scale mismatch outside margin m: $\mathcal{L}_{\text{mag}} = [|\log\rho_b| - m]_+$. The margin permits moderate differences in overall activation magnitude between modalities.

Latent-level modality balance. Matching aggregate magnitude does not ensure that corresponding phrases and regions activate the same latent features. For matched pair $(p, \pi(p))$, let $u_j = [s_{\pi(p)}]_j, \qquad v_j = [s_p]_j$ denote the visual and textual responses of latent j. We define its participation score as $\gamma_j = \text{sg}[\max(u_j, v_j)]$, where sg denotes stop-gradient, and retain the strongest participating latents $\mathcal{T}_p = \text{TopK}_{K_{\text{bal}}}(\gamma)$. Their normalized participation weights are $\omega_j = \frac{\gamma_j}{\sum_{j'\in\mathcal{T}_p}\gamma_{j'}}$, and the visual share of latent j is $\sigma_j = \frac{u_j}{u_j+v_j+\epsilon}$.

We penalize only excessive modality imbalance: $\mathcal{L}_{\text{bal}} = \frac{1}{|\mathcal{P}_b|}\sum_{p\in\mathcal{P}_b}\sum_{j\in\mathcal{T}_p}\omega_j\left[\left|\sigma_j - \frac{1}{2}\right| - \mu\right]_+^2$. The participation gate restricts this constraint to latents relevant to the matched content, while μ allows moderate modality specialization rather than forcing identical image and text responses.

Combined cross-modal objective. The complete phrase–region objective is $\mathcal{L}_{\text{cross}} = \lambda_{\text{al}}\mathcal{L}_{\text{align}} + \lambda_{\text{mag}}\mathcal{L}_{\text{mag}} + \lambda_{\text{bal}}\mathcal{L}_{\text{bal}}$. Together, these terms align phrase and region representations while maintaining compatible response scale and shared latent participation across modalities.

G.1.2 AuxK Activation for Dead-Latent Recovery

We use the AuxK objective of Gao et al. (2024) to encourage inactive latents to re-enter the SAE dictionary. AuxK reconstructs the residual error using features that are inactive under the main TopK pathway, providing an auxiliary gradient to otherwise unused latents. In our setting, however, the choice of activation used in this auxiliary pathway is important because many inactive latents have negative pre-activations.

Using a standard ReLU causes such latents to receive zero auxiliary activation and therefore no recovery signal. A direct alternative is to use the raw signed pre-activations, but we found this unstable: negative responses can contribute with unrestricted magnitude through the decoder and substantially overshoot the reconstruction residual. In our experiments, the resulting auxiliary reconstruction magnitude was approximately 7–14$\times$ larger than the residual it was intended to model, causing the auxiliary term to dominate the objective.

We also considered Softplus as a smooth alternative. However, $\text{softplus}(0) = \log 2$, so latents near the activation threshold contribute a nonzero constant even when their pre-activation is approximately zero. With a large dictionary, these small offsets accumulate through the decoder and introduce a systematic auxiliary reconstruction component. Softplus also alters the positive branch, making the auxiliary pathway inconsistent with the ReLU-based activation used by the main SAE.

We therefore use a leaky rectifier in the AuxK pathway,

$$\phi_{\text{aux}}(x) = \begin{cases} x, & x \geq 0, \\ \alpha x, & x < 0, \end{cases} \qquad \alpha = 0.01. \tag{8}$$

This preserves the standard ReLU response for positive pre-activations while providing a small nonzero gradient to negative inactive features. The modification therefore leaves the positive auxiliary path unchanged but allows persistently negative latents to receive a recovery signal.

As a sanity check, we constructed a synthetic setting in which a subset of candidate auxiliary latents remained negative throughout the forward pass. Under ReLU, these latents produced zero auxiliary

reconstruction and zero gradient. Replacing ReLU with the leaky rectifier yielded a finite auxiliary reconstruction and nonzero gradients for the same latents, confirming that the modification restores a recovery path without introducing the large responses observed with raw signed activations.

We use this leaky AuxK activation only in the auxiliary dead-latent recovery pathway; the main TopK encoding and reconstruction objective are unchanged.

H Implementation Details

SAE configuration. Across both backbones, we use an SAE with 8192 latents and TopK sparsity with $k = 128$. Representations are extracted from a single decoder layer, with placement matched approximately by relative depth across models: layer 15 of 28 for Qwen3-VL-2B-Instruct and layer 19 of 35 for Gemma-4-E2B-it, corresponding to relative depths of 0.536 and 0.543, respectively. Inputs are limited to 512 tokens, and images are resized to a maximum side length of 512 pixels.

Gemma intervention depth. Gemma-4-E2B-it shares key–value representations across its upper decoder layers: layers above the key–value boundary reuse keys and values computed earlier in the network. This architectural property does not affect experiments that intervene directly on all relevant token positions, but it prevents the patch-level causal grounding test from being evaluated at the primary $l = 19$ SAE placement.

Influence parameter scope. Gemma-4-E2B-it additionally contains an audio tower that is inactive for our image–text tasks. Parameters belonging exclusively to this unused modality receive no gradients and are therefore excluded from the influence-function parameter scope. All remaining active model parameters within the selected influence scope are included in the EK-FAC computation.

H.1 Warm-Start Initialization

We train the structured SAE in two stages. Stage A first learns a stable sparse dictionary using reconstruction alone, after which Stage B initializes from these learned SAE weights and introduces the additional shaping objectives. The second stage therefore begins from a well-formed dictionary rather than learning reconstruction, sparsity, localization, and cross-modal alignment simultaneously.

This warm start is important for optimization stability. Training Stage B directly from scratch with strong phrase–region alignment ($\lambda_{\text{al}} = 0.1$) causes substantial latent collapse, with 3,371 latents (41%) becoming inactive; adding load balancing does not prevent this failure. Increasing the amount of Stage-A pretraining progressively reduces collapse: 2,000 warm-up steps leave 1,927 dead latents (24%), while 6,000 steps reduce this to 142 (1.7%), close to the 56 dead latents obtained with the full Stage-A training. We therefore use the completed reconstruction-only Stage-A model to initialize Stage B throughout our experiments.

H.2 Dataset Processing and Task Construction

GQA. We construct GQA from the official balanced training and validation splits. To reduce redundancy from repeated questions about the same image, we retain at most one question per image using a fixed random seed. Training examples are sampled from `train_balanced`, while validation and test examples are disjoint subsets of `val_balanced`.

We use the original question verbatim as the prompt, without additional instructional scaffolding. This avoids introducing repeated boilerplate tokens that carry little example-specific information but can absorb attribution mass in latent and influence analyses. The target follows a *short-then-full* format: the canonical short answer appears first, followed by GQA's provided full-answer sentence. The short answer preserves the standard task output, while the full sentence provides sufficient linguistic structure for token-level multimodal attribution.

CheXpert. We use frontal chest radiographs and consider five findings: Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion. For each finding, labels are mapped to three classes: positive labels indicate *present*, uncertain labels remain *uncertain*, and all remaining labels are treated

as *absent*. We sample examples to balance these three classes independently for each finding, using a fixed random seed. Validation patients are held out from training to prevent patient-level leakage.

Each example is prompted with the image and the question *"For this chest X-ray, what is the status of {finding}?"*, with three answer choices corresponding to absent, uncertain, and present. The prediction target begins with the corresponding answer letter, followed by the cleaned report impression. We use this *impression-after* construction throughout the reported experiments, which preserves a concise classification target while providing clinically meaningful text for latent and attribution analysis.

VizWiz-Captions. We use the standard VizWiz caption annotations. We remove captions marked as rejected or precanned, as well as captions matching the dataset's precanned placeholder response. The prompt is fixed as *"Generate an accurate caption for this image."*

During model training, the target is a valid human reference caption. For latent necessity and sufficiency analyses, we instead use the finetuned model's own generated caption as the target. This evaluates whether perturbing a latent recovers or disrupts the model's actual prediction, rather than measuring agreement with an alternative human reference.

I Additional Modality Analysis

Section 4.4 studies how removing visual evidence changes training-example and latent influence rankings. Here we examine whether the strong image–text token-count imbalance itself biases the influence estimator.

I.1 Modality Imbalance

Vision–language sequences are strongly imbalanced in token count. Image tokens constitute approximately 78% of a CheXpert sequence and 93% of a VizWiz sequence. Because the influence computation aggregates over sequence positions, this raises a natural question: should image and text contributions be explicitly balanced? We study two possible corrections, acting on either the curvature estimate or the representation gradient.

Balancing the curvature estimate. A direct approach is to reweight token positions when estimating the EK-FAC factors so that image and text tokens contribute equal total mass. We interpolate between the standard and fully balanced estimators using

$$w_\alpha = (1-\alpha)\mathbf{1} + \alpha w, \qquad \alpha \in [0,1],$$

where w assigns modality-balanced token weights. The weighting is applied as $\sqrt{w_\alpha}$ to the corresponding EK-FAC factors so that the resulting covariance receives weight w_α.

This correction consistently degrades attribution quality. The CIDEr drop obtained by ablating the top 200 attributed latents decreases from 91.3% at $\alpha = 0$ to 90.4%, 85.6%, 80.4%, and 77.8% for $\alpha = 0.25, 0.5, 0.75$, and 1, respectively. Thus, equalizing modality mass in the curvature does not reveal a better operating point; performance decreases monotonically as the correction is strengthened.

This behavior is not caused by the curvature correction being negligible. The eigenvalue-corrected spectrum is highly concentrated, with a median-layer participation ratio of 0.0003 of the full dimension while 99.95% of entries remain nonzero. Reweighting therefore substantially changes the inverse-curvature operator, but emphasizes directions that are less useful for the downstream attribution objective.

Balancing the representation gradient. An alternative is to balance modalities after curvature correction, directly in the representation-space influence vector. Let s_{img} and s_{txt} denote the contributions aggregated over image and text positions. We form

$$\tilde{s}_\alpha = \alpha \frac{s_{\text{img}}}{\|s_{\text{img}}\|_2} + (1-\alpha)\frac{s_{\text{txt}}}{\|s_{\text{txt}}\|_2}, \tag{9}$$

and restore the magnitude of the unrestricted influence vector,

$$s_\alpha = \tilde{s}_\alpha \frac{\|s_{\text{all}}\|_2}{\|\tilde{s}_\alpha\|_2}. \tag{10}$$

Table 7: Influence with and without visual evidence. Latent ρ compares absolute latent influence, example ρ_{union} compares training examples active in either condition, and example ρ_{shared} is restricted to candidates retrieved in both conditions.

Dataset / Model	Pool J.	Latent ρ	Example ρ_{union}	Example ρ_{shared}	Top-8
GQA / Qwen	0.123	0.599	−0.564	0.260	0.073
GQA / Gemma	0.051	0.543	−0.755	0.183	0.000
VizWiz / Qwen	0.021	0.584	−0.793	0.279	0.010
VizWiz / Gemma	0.052	0.571	−0.774	−0.063	0.000
CheXpert / Qwen	0.145	0.737	−0.606	0.570	0.031
CheXpert / Gemma	0.026	0.795	−0.797	0.850	0.000

The final rescaling is important because influence depends on both alignment and magnitude: normalizing every example to unit norm would discard information about how strongly a training example affects the prediction.

In practice, this correction changes the ranking very little. On VizWiz with $\alpha = 0.5$, the balanced and standard influence rankings have Spearman correlation 0.994 over 60 evaluation examples. The top-ranked training example is unchanged for every example, and the mean top-8 overlap is 1.00.

Why modality balancing has little effect. The weak effect of gradient balancing follows from the structure of the learned latent space. Across latent dimensions, the image- and text-activation mass profiles have cosine similarity 0.73, or 0.733 when restricted to the $7{,}599$ of $8{,}192$ latents active in both modalities. Image and text therefore do not occupy approximately orthogonal subspaces; instead, they tend to activate many of the same latent directions in similar proportions.

Consistent with this observation, the cosine similarity between the modality-balanced vector and the original influence direction is 0.944, 0.988, 0.999, 0.973, and 0.916 for $\alpha = 0, 0.25, 0.5, 0.75$, and 1, respectively. In particular, the nominally balanced choice $\alpha = 0.5$ is almost identical to the unrestricted direction.

This also helps explain why curvature-side balancing is ineffective. Although image tokens dominate numerically, the imbalance does not correspond to a distinct "image direction" that can simply be downweighted relative to an independent text direction. Equalizing token mass therefore perturbs the curvature without isolating a complementary textual subspace.

Interpretation. These results indicate that token-count imbalance should not automatically be interpreted as bias in the influence estimate. The sequence is highly imbalanced at the input level, but the learned latent basis substantially overlaps across modalities. Consequently, simple reweighting either leaves the attribution nearly unchanged or degrades its causal fidelity. We therefore use the unweighted estimator in our main experiments and report modality balancing as a negative ablation.

I.2 Contrasting Influence With and Without Visual Evidence

We further analyze the image-removal experiment from Sec. 4.4. For each test example, influence is computed under matched conditions while changing only whether the image is present. The text-only condition therefore intervenes on the *input*, rather than modifying the learned latent dictionary. We compare latent influence correlation, candidate-pool overlap, example-level influence correlation over the union of active examples, and correlation restricted to shared candidates. We additionally report top-8 attribution overlap as an interpretable measure of ranking stability.

Across all six settings (Table 7), removing the image preserves substantially more agreement at the latent level ($\rho = 0.543$–0.795) than at the training-example level ($\rho = -0.797$–-0.564). Candidate pools also change substantially, with Jaccard overlap of only 0.021–0.145, while top-8 attribution overlap remains at or below 0.073.

Correlation among candidates retrieved in both conditions is more variable, ranging from -0.095 to 0.850. This suggests that the most consistent effect of image removal is a change in the *support* of training-data attribution—which examples are retrieved as influential—while subsequent re-ranking of shared examples depends more strongly on the model and task.

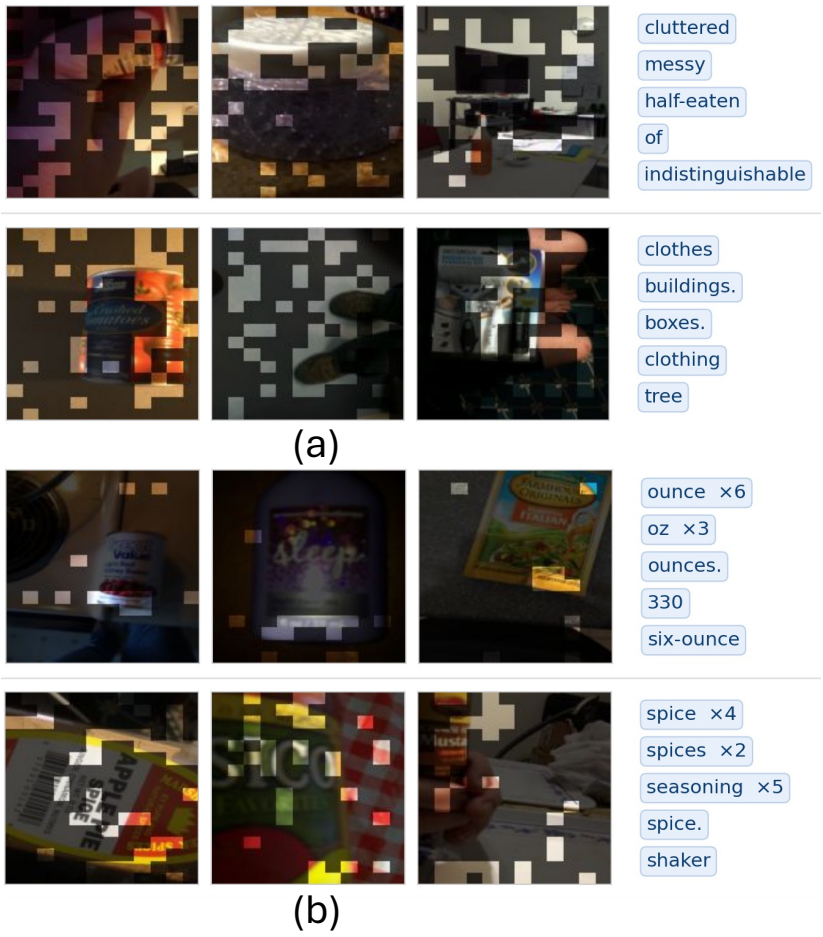


Figure 11: **Example latent comparison on VizWiz.** (a) Base SAE. (b) Our approach. Our latent exhibits a clearer semantic association and more spatially coherent visual support, while the base latent activates more diffusely across the image.

J Additional Qualitative and Hierarchical Attribution Analysis

We provide additional examples complementing Sec. 4.2 and Sec. 4.5. We separately examine test-conditioned latent decompositions and influential training examples to show how the attribution hierarchy behaves across examples and tasks. These examples complement the quantitative results in Table 3 by showing how the improved latent structure appears at the level of individual features. Across tasks, our objective produces latents that are more clearly associated with useful semantic terminology and whose visual support is more concentrated on the corresponding image evidence.

Figure 11 provides a corresponding comparison on VizWiz. The base SAE exhibits broader and less coherent visual support, whereas our objective produces latents with more interpretable semantic associations and activation concentrated on the image regions relevant to those semantics.

Finally, Fig. 12 shows representative GQA latents. Despite the different task structure, the learned features remain semantically interpretable and spatially concentrated on the image evidence associated with its textual meaning. Together, these examples illustrate that the improved localization and semantic quality observed quantitatively are also apparent across individual latents, datasets, and task domains.

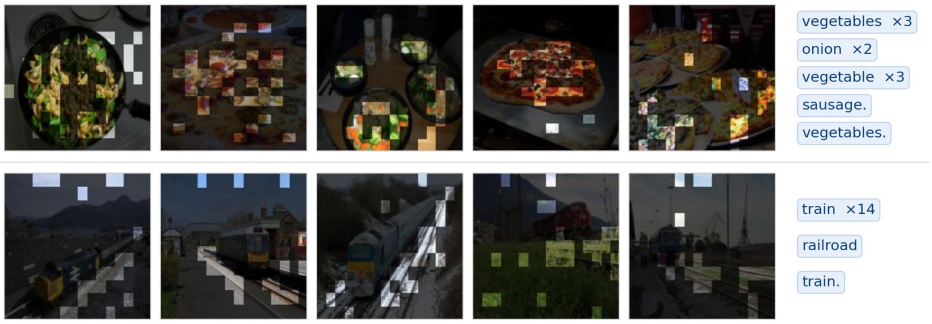


Figure 12: **Example latents on GQA.** Representative latents learned by our approach exhibits a meaningful semantic association together with localized activation over the corresponding visual evidence.

J.1 Detailed Analysis of Influential Test Latents

We further examine the latent decomposition of two individual VizWiz test examples to understand what information is represented beyond the aggregate latent-quality metrics. Rather than considering only the highest-scoring latent, we inspect a broader set of latents activated by each test image and compare their associated textual semantics with the image regions on which they respond. These examples show that the learned representation captures not only the primary object relevant to the prediction, but also complementary object parts, interactions, and scene context.

Figure 13 shows a test example centered on a medicine bottle. The highest-response latents recover several mutually consistent aspects of the scene, including concepts related to *bottle* and *medicine*. Importantly, other latents localize to the *hand holding the medicine*, capturing the interaction between the object and the person rather than simply firing on the bottle in isolation. This decomposition is particularly relevant to the test prediction: the model represents the object identity, its functional category, and the local visual evidence establishing how the object appears in the scene. The resulting latent set therefore provides a substantially richer account of the example than a single global visual feature.

A second example, shown in Fig. 14, contains a bottle of eye drops placed on a wooden table. Here, the latent decomposition again separates multiple visually and semantically meaningful components. Several latents correspond directly to the principal concepts needed to interpret the image, including *eyes*, *eye drops*, and *bottle*. At the same time, other latents capture contextual properties of the scene, such as the *wood* or wooden surface beneath the object. This contextual latent is notable because the test image itself contains a prominent wooden tabletop: although this evidence is not the primary answer concept, it faithfully reflects a real and localized component of the input rather than an unrelated semantic association.

Together, these examples highlight two useful properties of the learned latent space. First, semantically related concepts are decomposed across distinct but complementary latents rather than being collapsed into a single diffuse feature. Second, the associated visual support remains faithful to the specific test image: object latents localize to the relevant objects or parts, interaction latents capture relationships such as the hand holding the medicine, and contextual latents recover visible scene attributes such as the wooden tabletop. This fine-grained decomposition makes the latent representation useful not only for identifying *what* semantic evidence is present, but also for tracing *where* that evidence appears and how multiple pieces of evidence jointly support the model's prediction.

J.2 Qualitative Analysis of Influential Training Examples

We next inspect the training examples assigned the highest influence scores for individual test predictions. These examples provide a complementary view to the latent-level analysis: rather than asking which concepts within an example matter, we examine whether the retrieved training examples themselves contain evidence that is semantically consistent with the test query and prediction. Across VizWiz, GQA, and CheXpert, the most influential examples exhibit prediction-relevant

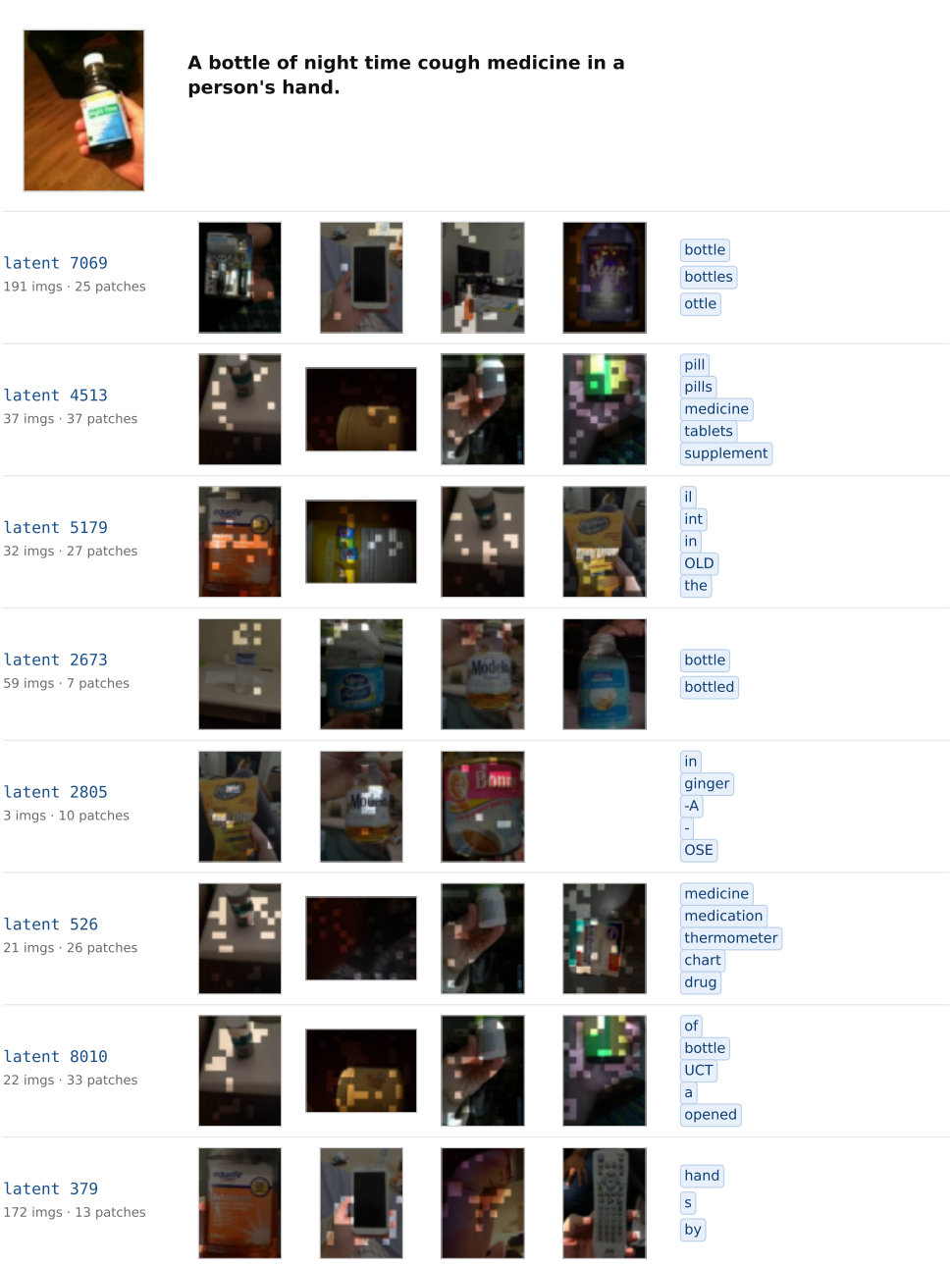


Figure 13: **Detailed latent decomposition for a VizWiz medicine example.** Representative test latents capture complementary evidence including the *bottle*, *medicine*-related semantics, and the *hand holding the medicine*. The localized activations show that the representation separates object identity, semantic function, and relevant object–person interaction while remaining grounded in the corresponding image regions.

correspondence with the test instance, ranging from direct visual or answer-level agreement to related clinical evidence across diagnostic labels.

We additionally expand the VizWiz medicine example introduced in the main text (Fig. 8) to inspect the influential training examples in greater detail. The highest-ranked examples capture several complementary aspects of the test instance rather than simply repeating the same object. Some

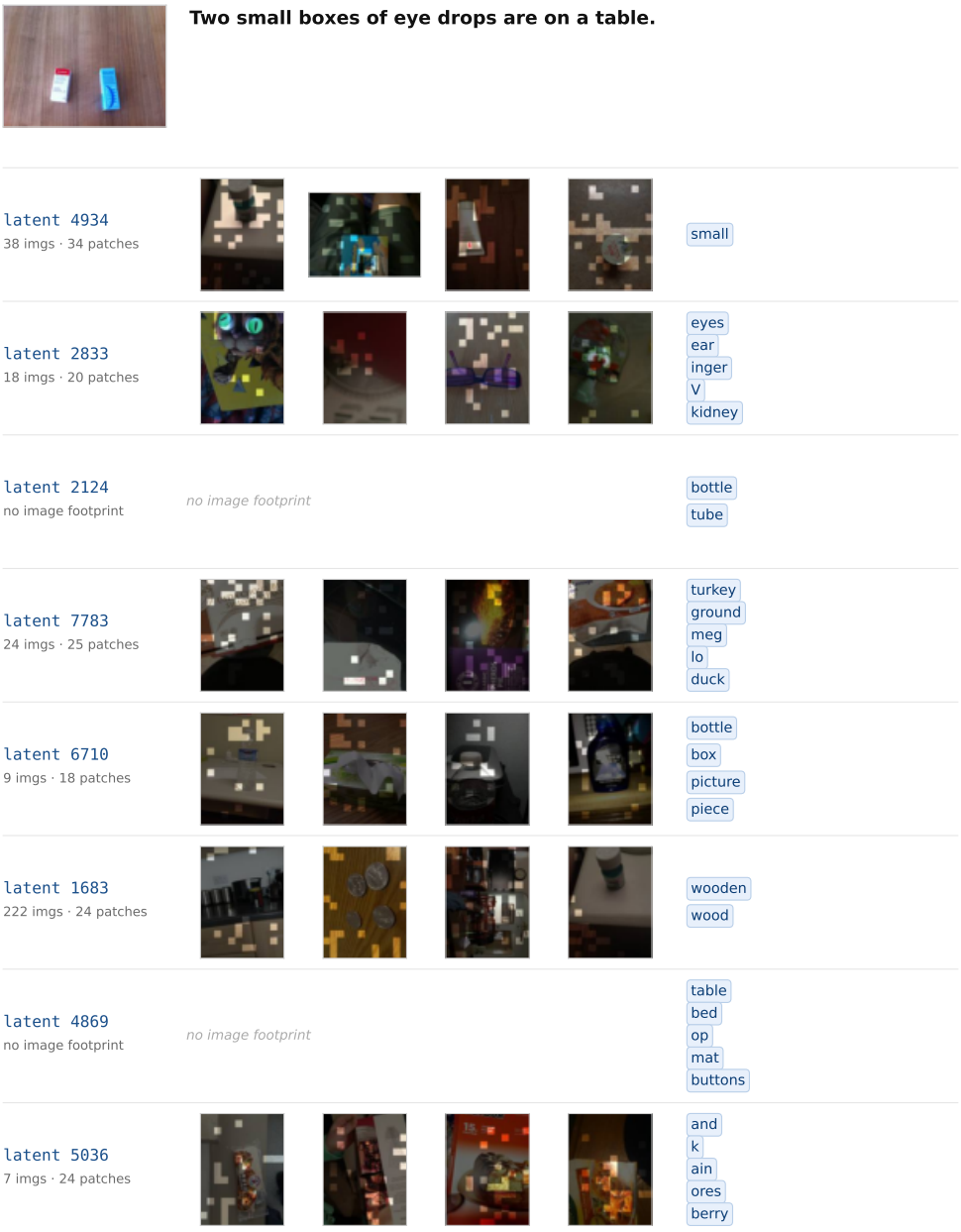


Figure 14: **Detailed latent decomposition for a VizWiz eye-drop example.** The learned latents recover the central semantics of the test image, including *eyes*, *eye drops*, and *bottle*, while also representing localized contextual evidence such as the *wooden table* on which the object appears. This example illustrates how distinct latents jointly encode both answer-relevant concepts and faithful scene context.

provide direct semantic support through examples involving *medicine* or *cough medicine*, while others contain visually related containers and packaging that help distinguish the target object from alternative possibilities. Additional examples reflect contextual properties present in the query image, including the wooden tabletop on which the medicine is shown. Together, these examples illustrate that the influence ranking captures both answer-relevant semantic evidence and visually

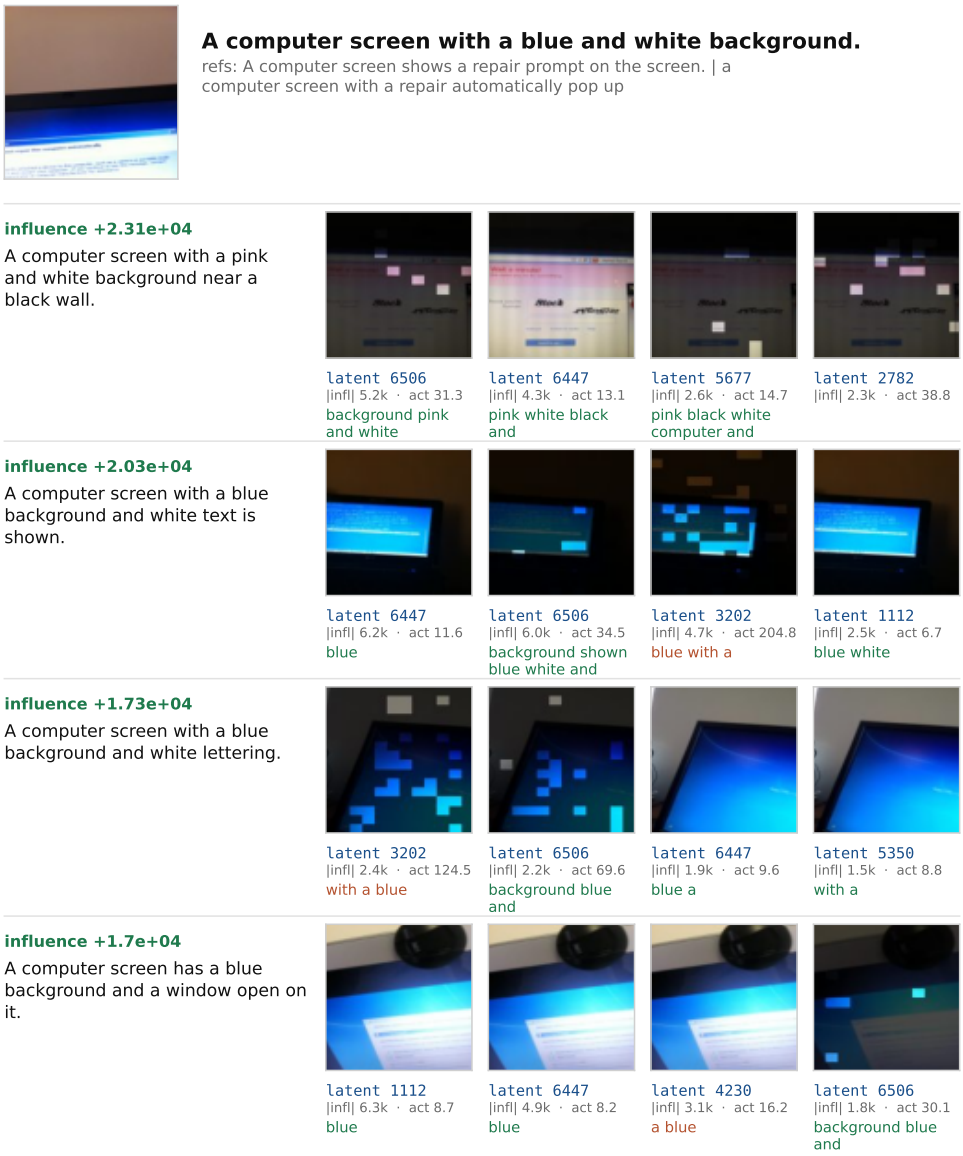


Figure 15: **Influential training examples for a VizWiz screen query.** The test example contains a screen, and the highest-influence training examples consistently exhibit closely related screen or display content. This qualitative consistency shows that influential examples reflect the salient visual semantics of the test input.

grounded contextual evidence, providing a more nuanced explanation of the prediction than simple nearest-neighbor similarity.

Figure 15 shows a VizWiz test example centered on a screen. The highest-ranked training examples consistently contain closely related screen or display content, indicating that the influence estimator retrieves examples with strong visual and semantic correspondence to the test input. Despite variation in the surrounding scene, the shared screen-related content remains prominent across the influential examples.

Figure 16 provides a GQA example in which the query asks about what a man is wearing. The model's answer is *jeans*, and the three most influential training examples are likewise associated

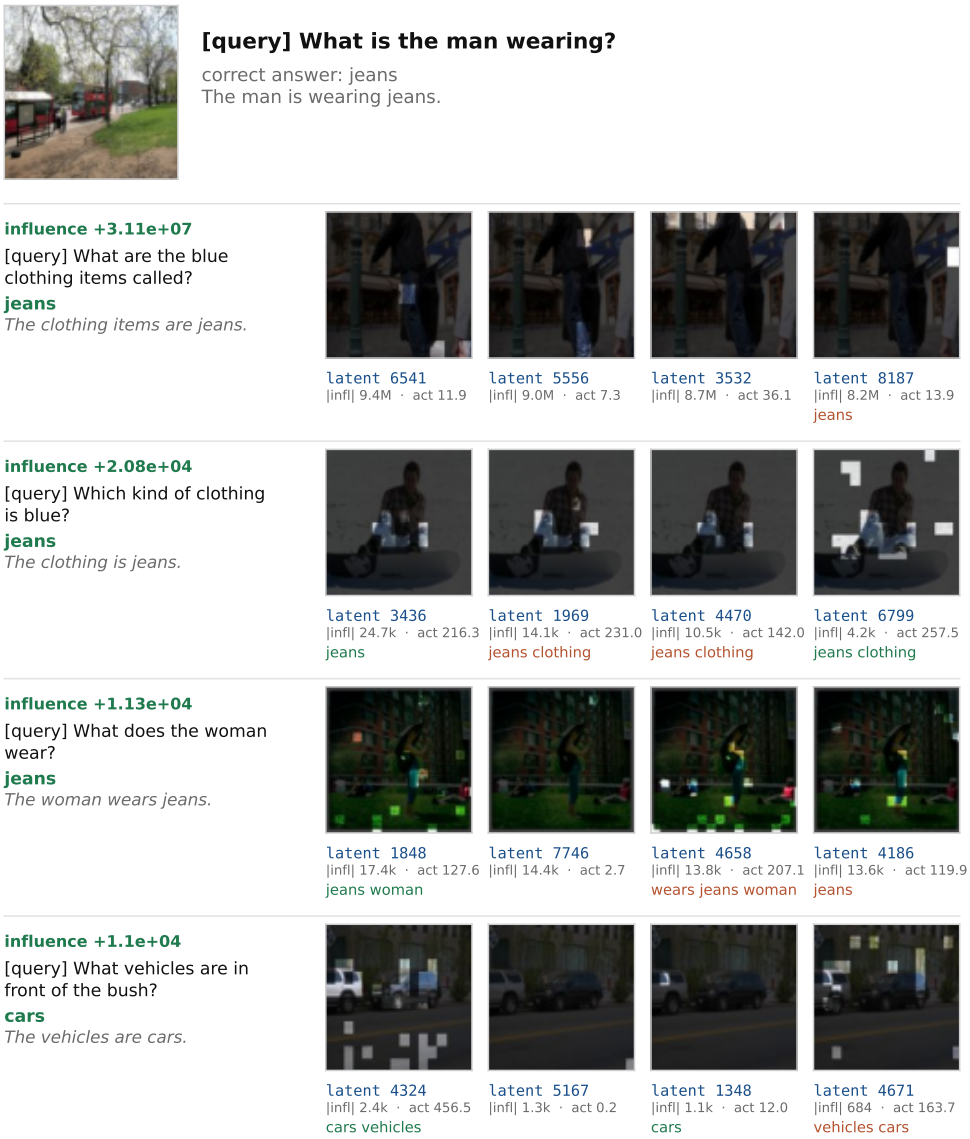


Figure 16: **Influential training examples for a GQA clothing query.** The test question asks what the man is wearing, with *jeans* as the predicted answer. The top three influential training examples are all associated with *jeans*, directly matching the answer semantics of the test query.

with *jeans*. The agreement is therefore not only visual but directly aligned with the answer concept, suggesting that the prediction is supported by training examples encoding the same attribute queried at test time.

A second GQA example is shown in Fig. 17. Here, the test question asks who is looking out the window, for which the answer is *man*. The highlighted answers of the most influential training examples are also *man*. This agreement indicates that the influence ranking identifies training examples whose supervision is directly consistent with the semantic decision made for the test instance.

We further examine influential training examples on CheXpert, where the prediction task differs from the open-ended settings above. Here, the highest-influence examples need not share the same

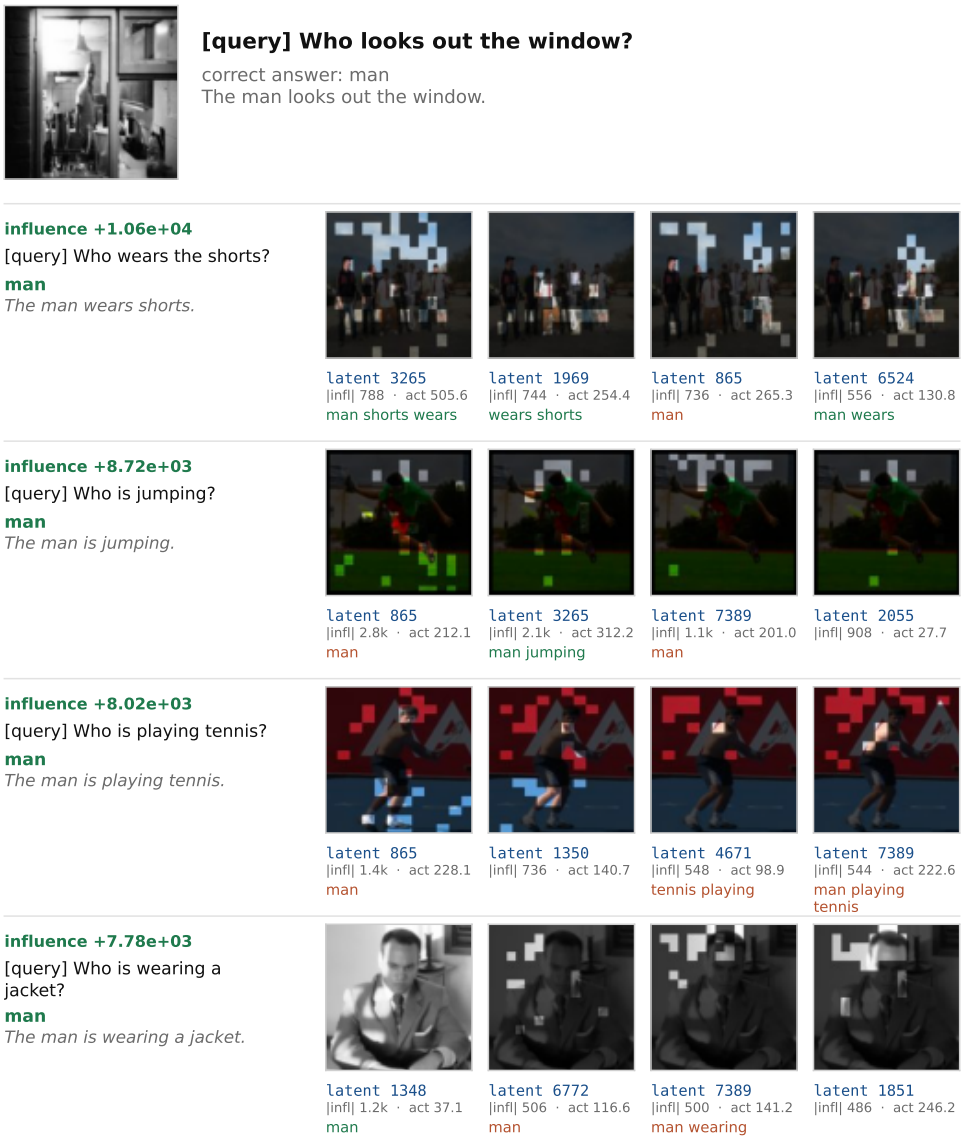


Figure 17: **Influential training examples for a GQA identity query.** The test question asks who is looking out the window, with *man* as the answer. The highlighted answers in the highest-influence training examples are likewise *man*, demonstrating strong agreement between the training supervision and the test prediction.

queried finding as the test example. Instead, they often contain related radiographic patterns or report semantics, providing a complementary view of how training examples support clinical predictions.

Figure 18 shows a correctly predicted *cardiomegaly* example. The influential training set spans several queried findings, including pleural effusion, consolidation, atelectasis, and cardiomegaly itself. Thus, the ranking is not explained simply by retrieving examples with the same diagnostic label. Rather, the influential examples capture overlapping cardiopulmonary evidence, while the localized latent activations identify the image regions through which this evidence contributes to the prediction.

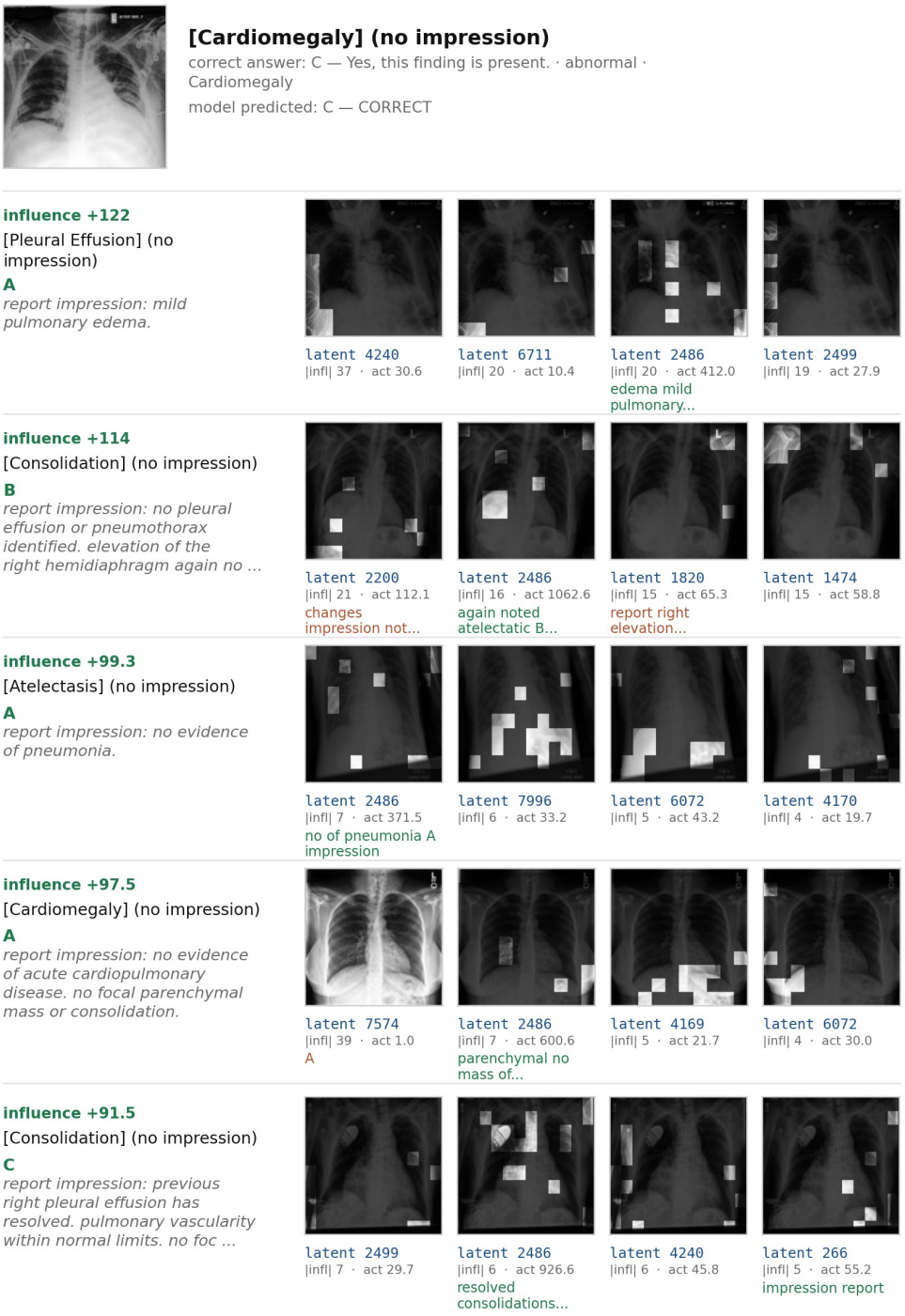


Figure 18: **Influential training examples for a CheXpert cardiomegaly prediction.** The correctly predicted cardiomegaly case receives support from training examples spanning multiple queried findings, including cardiomegaly and other cardiopulmonary abnormalities. This illustrates that influence is not determined solely by exact diagnostic-label matching.

The edema example in Fig. 19 makes this cross-label behavior particularly clear. Although the most influential training example is queried for *consolidation*, its report describes mild-to-moderate pulmonary vascular congestion, evidence that is semantically closely related to the edema prediction.

Table 8: Comparison with pooled multimodal SAE baselines on VizWiz with Qwen. Balance is computed over selective live latents. Footprint measures the fraction of the image patch grid activated by a latent, and contiguity is normalized against a size-matched random support. Lower footprint, contiguity, and word concentration indicate more selective features.

Method	Alive	Balance ↑	Footprint ↓	Contig. ↓	Word conc. ↓	Img. tok./fire ↓
MOSAIC	8156	**26.7**	**8.3%**	**0.548**	**0.219**	**3.30**
JSAE	99	3.3	100%	1.000	0.552	192
SPARC	69	0.0	100%	1.001	0.992	192

Other highly influential examples similarly contain pulmonary opacities and related abnormal chest findings. The retrieved examples therefore reflect the underlying evidence relevant to the prediction rather than merely the surface identity of the queried label.

A similar pattern appears for the *pleural effusion* prediction in Fig. 20. A directly matching pleural-effusion example appears among the highest-ranked training examples, while other influential cases are associated with cardiomegaly, atelectasis, and consolidation. The combination of direct and cross-label support suggests that the influence estimator captures shared radiographic evidence across related findings instead of reducing attribution to examples carrying the same task label.

Together, these cases show that high-influence training examples are not arbitrary nearest examples: they preserve prediction-relevant information from the test instance. The VizWiz example exhibits strong visual-semantic correspondence around the screen concept, while the GQA cases show direct agreement at the answer level for *jeans* and *man*. In CheXpert, the correspondence is more structured: influential examples may cross diagnostic labels while retaining related radiographic or report evidence. This provides qualitative evidence that the influence ranking surfaces training examples that are meaningfully aligned with the model's test-time decision rather than simply matching superficial labels or visual similarity.

K Comparison with Pooled Multimodal SAE Baselines

We compare MOSAIC against two recent multimodal SAE approaches that learn aligned image–text representations: JSAE (Shu et al., 2026) and SPARC (Nasiri-Sarvi et al., 2026). These methods provide natural controls for whether a generic shared or aligned multimodal latent space is sufficient for our attribution setting.

Experimental setup. All methods are trained on the same VizWiz activations from Qwen3-VL-2B-Instruct at layer 15. We use the same underlying VLM representations so that the comparison isolates the SAE objective. For SPARC, we use its published Global-TopK and cross-reconstruction formulation. For JSAE, we implement its separate image/text encoders and decoders with cosine alignment between the pooled modality codes. We use per-modality normalization and $\lambda_s = 1.0$, which places JSAE near its reported sparsity regime rather than evaluating it at the substantially denser operating point obtained with the original hyperparameter on our activations.

Global alignment does not imply localized shared features. Table 8 shows a substantial difference in the granularity of the learned representations. JSAE and SPARC can align image- and text-level representations, but their active visual features span essentially the full image. In contrast, MOSAIC retains thousands of live features while producing substantially more selective and spatially localized support. The distinction is particularly important for our use case: an example-level shared concept is insufficient to determine which image patch or text phrase carries a training example's influence.

This follows directly from the alignment objectives. JSAE aligns one mean-pooled image code with one mean-pooled text code for each example, while SPARC likewise learns correspondence at the representation-stream level. Neither objective defines a phrase–patch correspondence. In MOSAIC, the representation remains position-resolved and phrase–region supervision explicitly determines which localized visual and linguistic evidence should share latent structure.

Table 9: Effect of inserting each dictionary position-wise into the VLM forward path. Mean cross-entropy is evaluated over 60 VizWiz validation examples. Δ is relative to MOSAIC.

Forward path	Mean CE ↓	Δ
Identity (no SAE)	1.848	-0.180
MOSAIC	2.027	0
JSAE	9.464	$+7.44$
Mean-broadcast control	10.462	$+8.43$
SPARC	12.691	$+10.66$

Why pooled dictionaries cannot simply be applied per token. One possible alternative is to take a pooled multimodal dictionary and insert it at every image and text position, thereby creating the position-resolved representation required by latent influence. We test this directly by replacing the residual-stream representation with each dictionary's per-token reconstruction and measuring the resulting language-model loss.

As shown in Table 9, the pooled dictionaries do not preserve the VLM computation when applied position-wise at our intermediate layer. JSAE approaches the mean-broadcast control, while SPARC performs worse. This does not imply that these methods fail at their intended pooled representation or steering objectives; rather, their learned unit of analysis is not interchangeable with a token- or patch-level SAE inserted into the forward path.

This distinction matters for latent-mediated influence. Our attribution is computed through the SAE-mediated forward pass, so the prediction must depend on the same position-resolved latent variables that are subsequently attributed. A pooled representation can instead be analyzed in parallel, but then the model loss does not depend on those codes and their gradients cannot mediate the prediction. MOSAIC is therefore designed specifically to preserve the forward computation while providing a shared latent basis at the phrase, token, patch, and region levels needed for grounded influence decomposition.

Takeaway. These comparisons isolate the difference between *cross-modal alignment* and an *attribution-compatible multimodal basis*. Existing pooled multimodal SAEs can learn shared image–text concepts, but do not natively provide the position-resolved, spatially localized representation required to trace training-example influence to specific linguistic and visual evidence.

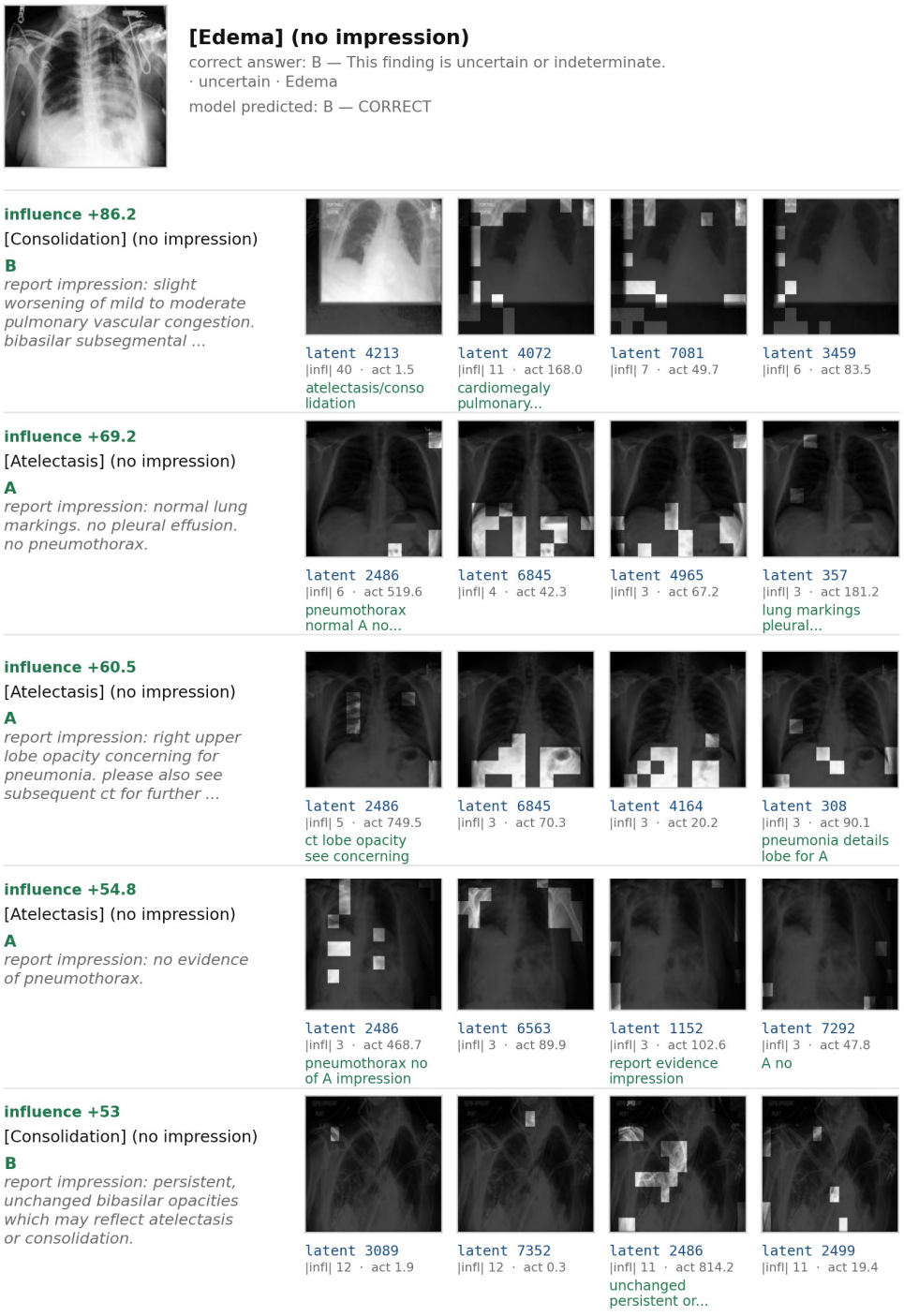


Figure 19: **Influential training examples for a CheXpert edema prediction.** The highest-influence example is queried for consolidation but its report describes pulmonary vascular congestion, providing evidence closely related to edema. The ranking therefore captures clinically relevant cross-label correspondence rather than only exact label agreement.



[Pleural Effusion] (no impression)

correct answer: C — Yes, this finding is present. · abnormal · Pleural Effusion

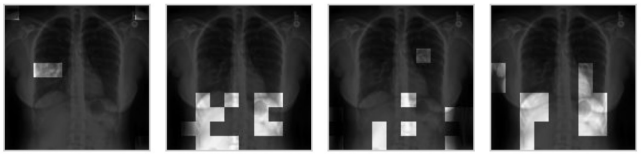
model predicted: C — CORRECT

influence +83.2

[Cardiomegaly] (no impression)

A

report impression: right middle lobar patchy opacity. findings most suggestive of infectious pneumonia, less likely ...



latent 2486 | |infl| 4 · act 756.6 | patchy infectious lobar opacity...

latent 6845 | |infl| 4 · act 39.5

latent 4965 | |infl| 4 · act 50.2

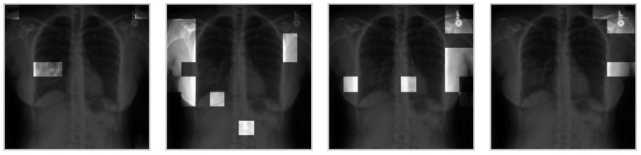
latent 7996 | |infl| 3 · act 57.3

influence +81.5

[Atelectasis] (no impression)

A

report impression: right middle lobar patchy opacity. findings most suggestive of infectious pneumonia, less likely ...



latent 2486 | |infl| 4 · act 761.2 | patchy lobar infectious...

latent 6563 | |infl| 4 · act 117.6

latent 1392 | |infl| 4 · act 42.2

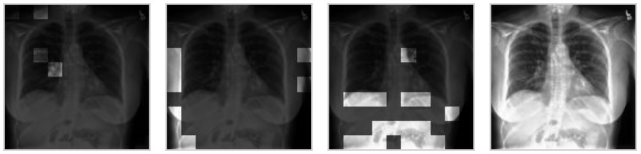
latent 1820 | |infl| 3 · act 113.5 | opacity report pneumonia righ...

influence +76.2

[Pleural Effusion] (no impression)

A

report impression: chest 2 views, demonstrate no focal consolidation or pleural effusion. low volume. cardia ...



latent 2486 | |infl| 5 · act 764.6 | 2 views vascularity...

latent 2499 | |infl| 3 · act 39.8

latent 7996 | |infl| 3 · act 58.9

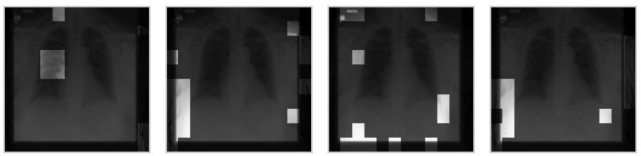
latent 6731 | |infl| 3 · act 2.8 | demonstrate

influence +65.1

[Consolidation] (no impression)

A

report impression: chest 1 view, demonstrate no focal consolidation or pleural effusion. low volume. cardia ...



latent 2486 | |infl| 8 · act 794.6 | view vascularity 1 consolidation A

latent 2499 | |infl| 7 · act 56.8

latent 1043 | |infl| 5 · act 65.8

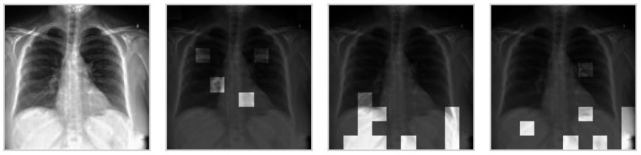
latent 7065 | |infl| 5 · act 13.1

influence +64.8

[Cardiomegaly] (no impression)

A

report impression: frontal and lateral views of the chest demonstrate a normal cardiomediastinal silhouette ...



latent 7574 | |infl| 24 · act 1.0 | A

latent 2486 | |infl| 4 · act 1006.1 | focal clear A of impression

latent 4169 | |infl| 3 · act 20.8

latent 2200 | |infl| 2 · act 69.2 | impression unremarkable...

Figure 20: **Influential training examples for a CheXpert pleural effusion prediction.** The influential set contains both a directly matching pleural-effusion example and examples associated with other thoracic findings, demonstrating a mixture of direct and cross-label support for the prediction.