FROM ACTIVATION TO INTERVENTION: INTERPRETING SPARSE VISUAL FEATURES THROUGH RECONSTRUCTION

Anonymous authors
Paper under double-blind review

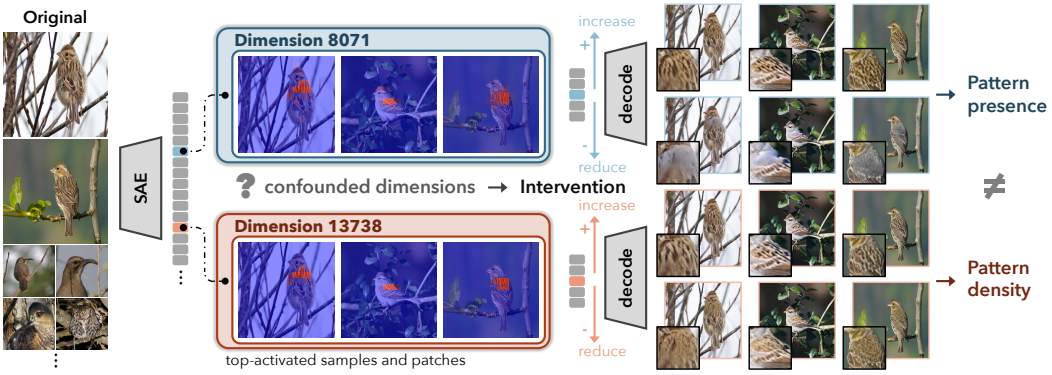


Figure 1: **Fine-grained categories expose confounded SAE dimensions**. Top-activated samples and pixel regions cannot effectively interpret the semantics they represent. We apply **dimensional intervention** on patch embeddings and decode them back to images. Comparing the visual changes from reducing or increasing the corresponding activation more clearly reveals the patterns or semantics represented by various SAE dimensions.

ABSTRACT

Sparse autoencoders (SAEs) decompose visual representations into individual features, offering a promising basis for explaining model behavior. Interpreting these features typically relies on images or pixel regions that most strongly activate them. However, this evidence becomes ambiguous when multiple features with different semantics are frequently activated in the same locations, such as texture and color of the same patterns. We propose **Reconstructive SAE Intervention** (**RECSAEI**), a framework that shifts feature explanation from passive observation to active visual reconstruction. By intervening on representations along SAE decoder directions and reconstructing the visual output, we provide direct evidence of a feature's semantic meaning. We demonstrate that our method successfully disambiguates feature meanings across multiple datasets and vision encoders. Furthermore, our framework grounds `[CLS]` SAE interventions in a sparse set of patch SAE decoder directions. Beyond stable target response, this attribution also exposes the compositional relationship between elementary visual patterns and high-level semantics. Our results advocate for intervention reconstruction as an essential complementary signal for explaining neural networks.

1 INTRODUCTION

Vision transformers process visual evidence across patch and `[CLS]` tokens to produce downstream predictions (Dosovitskiy et al., 2021). Identifying the precise evidence that supports these predictions, however, is difficult within dense representations (Elhage et al., 2022; Bau et al., 2017). Sparse autoencoders (SAEs) address this challenge by decomposing dense representations into individual sparse features, each corresponding to a dimension of the sparse representation (Shu et al., 2025; Huben et al., 2024; Fry, 2024; Daujotas, 2024a;b). Understanding what these individual features represent is then essential for building meaningful explanations of model behavior.

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.

For vision models, a common practice to interpret an SAE feature is to retrieve and visualize the components (either regions within a given image or images from a given dataset) that maximally activate it (Rao et al., 2024; Lim et al., 2025; Stevens et al., 2025). Such activation-prioritized visualizations suggest possible interpretations, but several visual properties may consistently appear together, leaving the meaning of an individual feature ambiguous. This problem is exacerbated in fine-grained scenarios (Wei et al., 2021): as shown in Figure 1, two SAE dimensions can activate on nearly identical images and regions while representing different properties. Their shared activation patterns reveal where relevant evidence occurs, yet leave unclear which aspect of that evidence each feature encodes. Intervention approaches offer additional evidence by measuring how modifying a feature changes downstream predictions (Chatzoudis et al., 2025; Goyal et al., 2019a; Zhou et al., 2018; Abid et al., 2022; Bau et al., 2020). Yet a change in a category prediction alone does not identify the visual property associated with the intervened feature. Resolving this ambiguity requires making the visual consequences of feature intervention directly observable.

We propose **Reconstructive SAE Intervention**, or **RECSAEI**, which couples **sparse-feature intervention** with **reconstruction** of the modified representation into an image. We first apply this approach to SAEs trained on patch tokens. Each target feature defines a direction in the patch representation through its SAE decoder vector. We modify the original patch embeddings along this direction and reconstruct the resulting image using a representation autoencoder (RAE) (Zheng et al., 2026; Singh et al., 2026). Comparing reconstructions across intervention strengths **reveals the visual changes associated with the target feature**. As shown in Figure 1, these changes distinguish the pattern presence and density, despite their similar high-activation examples.

Patch-level intervention, however, often does not directly connect to downstream predictions. Therefore, we also need to interpret `[CLS]` SAE features, which integrate information from patch representations and contribute more closely to model behavior. Yet the RAE decoder reconstructs images from patch representations and cannot directly decode the `[CLS]` token. We therefore connect the target `[CLS]` SAE feature **back to the patch SAE features through gradient attribution**. We identify patch features that are both active and influential for the target, then combine their SAE decoder vectors into an intervention direction. Modifying patch representations along this direction changes the target `[CLS]` activation and enables reconstructing the associated visual change.

Compared with direct steering along a dense gradient, RECSAEI produces **more stable visual change and consistent responses** in the target `[CLS]` feature (§4.2). The sparse attribution also **identifies the patch SAE features** that support each target `[CLS]` feature (§4.3). Our analysis shows that these supporting sparse components may have distinct activation patterns and intervention effects. Their joint contribution nevertheless supports a stable and coherent higher-level response. More broadly, our results motivate shifting feature explanation *from activation patterns alone to examining the visual changes they produce under intervention*. Such intervention-based explanations provide more precise evidence for feature semantics and trace how patch-level patterns support image-level concepts, opening a path toward richer mechanistic analyses of vision models.

2 RELATED WORK

Sparse Concepts in Visual Representations. Understanding concepts encoded in visual representations helps reveal the evidence underlying model decisions. Existing methods relate representations to concepts through individual units (Bau et al., 2017), predefined concept directions (Kim et al., 2018), concept bottlenecks (Koh et al., 2020), clustering (Ghorbani et al., 2019), or factorization (Zhang et al., 2021; Fel et al., 2023). Sparse dictionary learning provides a complementary direction by representing activations with a small number of elements from an overcomplete dictionary (Olshausen & Field, 1996; Aharon et al., 2006; Donhauser et al., 2025). Sparse autoencoders (SAEs) extend this direction with an encoder-decoder reconstruction objective (Fry, 2024). In vision, SAEs have been applied to image-level embeddings (Fry, 2024; Daujotas, 2024a;b; Rao et al., 2024) and multimodal representations (Shen et al., 2026). More recent work applies SAEs to patch embeddings to expose spatially localized features (Lim et al., 2025; Stevens et al., 2025), evaluates their monosemanticity (Pach et al., 2025), and identifies encoded concepts such as visual biases (Choi et al., 2026).

A central challenge for concept-based methods is establishing what an individual concept represents. Existing SAE studies commonly interpret each dimension from its highest-activating examples (Zhang et al., 2025a; Pahuja et al., 2026). Ferrando et al. (2026) use language models to summarize the

effects of steering sparse features in visual embeddings, but rely on the aligned multimodal space of language models. Other work combines sparse features with explicitly interpretable concept spaces to improve robustness of interpretation (Kulkarni et al., 2026). Together, these methods provide rich descriptions of sparse features, but establishing their precise visual semantics remains challenging.

Intervention of visual concepts. Many visual interpretability methods infer relevant evidence from activations, gradients, or input perturbations (Selvaraju et al., 2017; Wang et al., 2020; Chefer et al., 2021; Fong & Vedaldi, 2017; Goyal et al., 2019b; Zhang et al., 2025b; Chowdhury et al., 2025). Intervening further confirms the functions of that evidence by revealing changes in visual effects or predictions. Concept bottleneck models allow predefined concepts to be manipulated and propagate their effects to the output (Koh et al., 2020; Shin et al., 2023; Steinmann et al., 2024; Abid et al., 2022). However, they cannot reveal the semantics of internal model representations. Prior work has also manipulated generative models for semantic image editing (Goyal et al., 2019a; Shen et al., 2020; Härkönen et al., 2020; Shen & Zhou, 2021). Recent diffusion methods realize fine-grained generation control (Hertz et al., 2023; Tumanyan et al., 2023; Brack et al., 2023; Gandikota et al., 2024). However, plugging concepts from discriminative models into generative models often needs dedicated training for each representation type. The iterative denoising process also adds uncertainty in the faithfulness of the intervened feature.

SAE provides another form of direct latent intervention. Manipulating the decomposed sparse features can mitigate object hallucination (Park et al., 2026), improve zero-shot classification (Chatzoudis et al., 2025), and steer diffusion generation (Kim & Ghadiyaram, 2025). Carlyn et al. (2026) further use sparse latent interventions to produce localized counterfactual edits. These methods primarily use intervention to control model behavior or generation. Kulkarni et al. (2026) steer SAE dimensions to explain different features, but rely on language models. We instead use reconstruction to recover the visual change and use it as evidence for disambiguating the semantics of the sparse feature.

3 METHODOLOGY

Assume an SAE model trained on the internal representations of a pretrained vision encoder, *e.g.*, DINOv3 (Siméoni et al., 2026). Interpretability methods can identify the correlation between specific sparse dimensions and downstream predictions (Steinmann et al., 2024; Chatzoudis et al., 2025). We then need to understand the semantics of the visual evidence to fully explain model behavior. For a sparse dimension k, a common way to infer its semantics is to inspect examples and spatial locations with large activation (Zhang et al., 2025a; Pahuja et al., 2026). Such evidence describes the inputs associated with a feature, but does not isolate the exact semantics. In fine-grained representations, in particular, the same image pixels may contain several co-occurring properties, such as the presence/absence and density of a pattern that can be used to differentiate categories or variants. High-activating examples alone cannot reliably distinguish such ambiguity.

Intervention provides complementary functional evidence by directly modifying a feature and observing its downstream effect (Shin et al., 2023; Steinmann et al., 2024). However, the connection between a sparse dimension k and prediction still does not specify the represented property. To this end, we propose **Reconstructive SAE Intervention** (RECSAEI), which couples sparse-feature intervention with *reconstruction* to decode the intervention back into image space and inspect the corresponding visual change. Specifically, we use a pretrained RAEv2 representation decoder (Singh et al., 2026). Given patch representations from the corresponding vision encoder, the decoder reconstructs an image. Decoding intervened embeddings thereby exposes the visual consequence of the intervention. In this way, representation reconstruction enables translation between feature intervention and appearance changes, offering a mechanism to probe sparse visual features beyond activation patterns. The question, then, is *how we can derive a reasonable intervention that can be reliably decoded*. We start with intervening on a single patch SAE dimension.

3.1 INTERVENTION ON A PATCH SAE DIMENSION

Let $\boldsymbol{h}_p \in \mathbb{R}^d$ denote the encoded representation of a spatial patch token p, $\boldsymbol{z}_p = E_{\text{SAE}}(\boldsymbol{h}_p)$ its sparse feature activations, and $D_{\text{SAE}}(\boldsymbol{z}_p) = \boldsymbol{W}\boldsymbol{z}_p + \boldsymbol{b}$ its SAE reconstruction. We intervene on the sparse dimension k by scaling its activation as $z'_{p,k} = \gamma z_{p,k}$, with the other dimensions unchanged. A

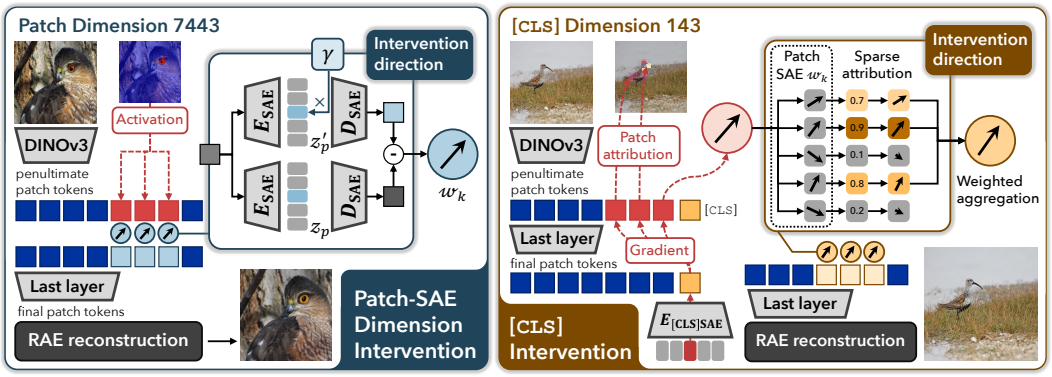


Figure 2: **Overview of RECSAEI for patch and [CLS] SAE dimensions**. **Left**: For a patch-SAE feature, we identify highly activated spatial patches and intervene on the corresponding sparse dimension along the decoder direction $\boldsymbol{w}_k$. **Right**: For a [CLS] feature, we first attribute its activation to patch tokens through backpropagation. After decomposing the patch-level gradient into patch-SAE dimensions, we aggregate top-K decoder directions $\boldsymbol{w}_k$ to obtain the intervention direction. We apply the intervention to attributed patches and decode changes through RAE reconstruction.

naive approach would decode the modified sparse representation directly through $\boldsymbol{h}_p^{'} = D_{\text{SAE}}(\boldsymbol{z}_p^{'})$. However, SAE reconstruction may introduce errors across all dimensions that conflate the intended feature intervention. Therefore, as shown in the left panel of Figure 2, we retain the original patch tokens and add only the change induced by the selected feature:

$$\boldsymbol{h}_p^{'} = \boldsymbol{h}_p + \left[D_{\text{SAE}}(\boldsymbol{z}_p^{'}) - D_{\text{SAE}}(\boldsymbol{z}_p)\right] = \boldsymbol{h}_p + \boldsymbol{W}(\boldsymbol{z}_p^{'} - \boldsymbol{z}_p) = \boldsymbol{h}_p + (\gamma - 1) z_{p,k} \boldsymbol{w}_k,$$

where $\boldsymbol{w}_k$, the k-th column of $\boldsymbol{W}$, is the decoder direction associated with sparse dimension k. The intervention inherently moves $\boldsymbol{h}_p$ along the feature's direction in representation space. This formulation recovers the original representation when $\gamma = 1$. It also enables further extrapolation in reducing the feature activation below 0.

Applying intervention across all spatial patches may induce unintended changes in image regions that contribute less to the feature's representation. Therefore, we restrict each intervention to its dominant spatial support. Specifically, we rank patches by feature activation and retain the smallest set that accounts for at least 80% of the total activation mass, with at least four patches. The excluded patches are left unchanged. We apply the interventions on the input patch embeddings of the last layer and propagate them through the remaining encoder computation, after which they are decoded into full images using the RAE decoder (Singh et al., 2026).

3.2 INTERVENTION ON [CLS] FEATURES

Whereas patch-level SAE features are defined on individual spatial token representations, [CLS] features describe more complex information, integrated from preceding patch tokens (Dosovitskiy et al., 2021). Intervening on these features reveals higher-level semantics and connects them to downstream predictions. However, the RAE decoder reconstructs images from patch representations. Directly modifying the final [CLS] token therefore does not produce a corresponding image intervention. To this end, we propose to attribute the target [CLS] feature back to the patch tokens and transform this signal into a sparse combination of patch-SAE decoder directions. Steering the selected patch representations along the aggregated direction and propagating the intervention through transformer blocks will alter the target [CLS] feature and produce a corresponding visual reconstruction.

Sparse attribution. Assume a target SAE trained on final-layer [CLS] representations and a source SAE trained on penultimate-layer patch representations. Let y_c denote the activation of target [CLS] SAE feature c. We backpropagate the activation to derive patch-level gradients $\boldsymbol{g}_p = \nabla_{\boldsymbol{h}_p} y_c$, and localize the intervention by selecting the smallest set of patches $\mathcal{P}$ accounting for at least 80% of the total gradient-norm mass. We then express the target's sensitivity in the source SAE dictionary. Let $z_{p,k}$ be the source activation of feature k at patch p, and $\boldsymbol{w}_k \in \mathbb{R}^d$ its decoder vector. The directional sensitivity of the target to this feature is $q_{p,k} = \boldsymbol{g}_p^\top \boldsymbol{w}_k$. To identify features that are both present in

the image and influential for the target, we aggregate absolute activation-weighted sensitivities for each source dimension over the selected patches:

$$\sigma_k = \sum_{p \in \mathcal{P}} \left| z_{p,k} q_{p,k} \right|.$$

We retain the K highest-scoring features with nonzero attribution, forming a set $\mathcal{S}_K$. If fewer than K features have nonzero scores, we retain only those features. Absolute values prevent positive and negative contributions at different locations from cancelling during feature selection.

Steering aggregation. Using the selected source features, we construct a spatially varying steering direction by combining their decoder vectors with signed directional sensitivity weights:

$$\boldsymbol{v}_p = \sum_{k \in \mathcal{S}_K} q_{p,k} \boldsymbol{w}_k, \qquad \boldsymbol{u}_p = \frac{\boldsymbol{v}_p}{\sqrt{\sum_{p \in \mathcal{P}} \|\boldsymbol{v}_p\|_2^2}}.$$

Here, the signed sensitivity $q_{p,k}$ determines both the magnitude and direction of each source feature's contribution. The normalization fixes the overall steering-direction norm across the selected patches $\mathcal{P}$. We then apply the intervention to the selected patches:

$$\boldsymbol{h}_p^{'} = \boldsymbol{h}_p + s \sum_{p \in \mathcal{P}} \|\boldsymbol{h}_p\| \boldsymbol{u}_p,$$

where s controls the intervention strength. Scaling by the aggregate norm of the original selected patch representations makes s a relative intervention magnitude and provides more consistent control across images. We pass the edited patches through the final encoder block and reconstruct the image using the RAE decoder. The semantics revealed by observing visual changes after reconstruction can then be incorporated with other evidence to explain model behavior more accurately.

4 Experiments

4.1 Intervention as Explanation

We first evaluate whether the intervened reconstruction provides meaningful information for explaining patch sparse features. We perform the study on four datasets: NABirds (Van Horn et al., 2015), CUB (Wah et al., 2011), Cars (Krause et al., 2013), and ImageNet (Deng et al., 2009). In fine-grained scenarios, sparse features represent subtler semantics than in coarse-grained datasets. CLIP (Radford et al., 2021) or GroundingDINO (Liu et al., 2024) models may not accurately identify fine-grained organisms. Therefore, IoU or CLIP score adopted in previous literature (Xu et al., 2025; Zhang et al., 2025a) cannot faithfully reflect explanation quality (results in §C.2). Moreover, many features are hard for humans to describe and difficult to match to a predefined dictionary.

To this end, we propose to predict feature activation on held-out natural images. Specifically, for each sparse feature, we construct its explanation using the training set. Given two test images and the explanation, an evaluator predicts which image has a higher activation for this feature.

Experiment setting. We build a paired benchmark for this evaluation. First, we exclude dense (activated on more than 0.1% of images) and dead dimensions. Within the remaining dimensions, we exclude dimensions with more than 20% activation on the backgrounds. We use bounding-box labels from NABirds, CUB, and Cars to indicate foreground, and generate bounding boxes for ImageNet using GroundingDINO and the category name. From the resulting candidate pool, we randomly select $1{,}000$ dimensions to build the benchmark. For each dimension, we select five image pairs with high cosine similarity in the dense embedding space but exhibit a strong contrast in SAE activation. §B and Table 4 present more details of the benchmark. We use GPT-5.6-Terra (OpenAI, 2026) to predict the sample with higher activation values and compare the following forms of explanations: (1) text explanations generated by observing the top examples, *i.e.*, images with high activations in the target feature; (2) top examples; (3) top examples with intervention reconstructions; (4) top examples and bottom examples, *i.e.*, images with low activations; (5) the combination of top examples, intervened reconstructions, and bottom examples. We provide the activation overlay along with top examples to indicate feature location (Ablation in §C.1). We include embeddings from DINOv3 (Siméoni et al., 2026) and SigLIP 2 (Tschannen et al., 2025). Training and intervention details are included in §A.

Table 1: **Feature-explanation accuracy using different evidence**. Text: descriptions of the target feature. Top: images and regions with high activations. Bottom: images with low activations. Intervention: reconstructions after reducing and increasing the feature activation. Intervention reconstructions provide information complementary to both top and bottom examples to explain SAE features.

Explanation approach				NABirds		CUB		Cars		ImageNet	
Text	Top	Bottom	Intervention	DINOv3	SigLIP 2	DINOv3	SigLIP 2	DINOv3	SigLIP 2	DINOv3	SigLIP 2
✓				62.8	59.8	59.0	59.1	61.5	61.4	74.1	71.1
	✓			66.3	61.0	64.3	63.1	67.4	65.5	74.7	71.8
	✓		✓	68.2	63.8	65.4	64.8	68.4	66.0	71.4	72.3
	✓	✓		73.0	67.9	70.3	69.7	73.2	69.4	**76.9**	73.5
	✓	✓	✓	**74.8**	**69.2**	**71.4**	**71.1**	**74.3**	**71.0**	76.2	**73.9**

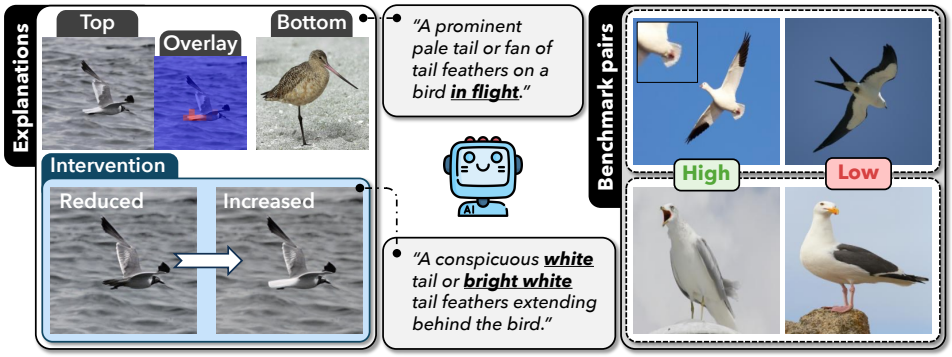


Figure 3: **Qualitative example of intervention-assisted feature explanation**. Top and bottom examples alone might bias the benchmark prediction toward pose, prominence, or "in-flight" status. Intervention reconstruction isolates the more specific change toward bright white tail feathers, leading to a more precise feature interpretation.

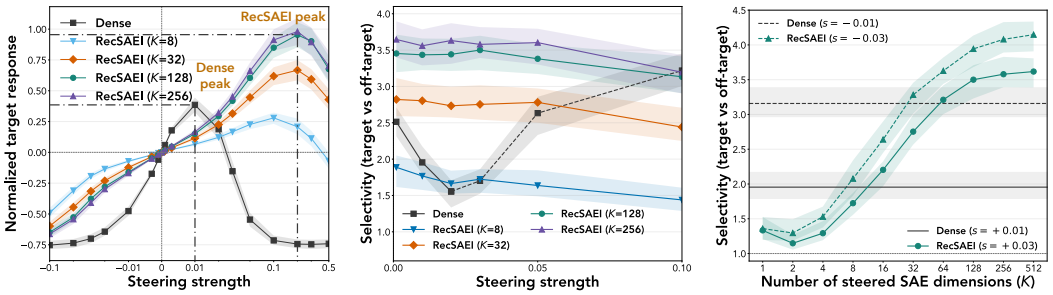
Results. We present the accuracy of predicting the sample with higher activations in Table 1. Noticeably, top-activating examples already outperform text-only explanations across all settings, even though the descriptions are produced by the same GPT-5.6-Terra model. Directly comparing the test images with examples facilitates more flexible and comprehensive reasoning. Bottom-activating examples can be treated as indirect intervention on the target dimension, which provides a substantial performance gain, suggesting that contrasting when a feature is present and absent is highly informative. However, bottom examples cannot directly disambiguate confounded features.

Intervention reconstruction (RecSAEI) contributes a further, largely complementary signal. When added to top and bottom examples, it improves explanation accuracy in seven of eight settings. The improvement is more significant in fine-grained datasets, including gains of 1.8 points on NABirds with DINOv3 and 1.6 points on Cars with SigLIP 2. ImageNet contains coarser-grained categories, making it easier to identify the higher-activating test image. Nevertheless, intervention reconstruction still improves the explanation accuracy for SigLIP 2. Overall, the results suggest that feature activation and intervention characterize different aspects of a sparse dimension: activation indicates the visual contexts associated with the feature, while intervention helps isolate the visual property controlled by that feature. Observing how a feature changes visual appearance provides semantic evidence that its activation pattern alone cannot recover. We also evaluate text descriptions produced with different visual components in §C.3 to avoid bias from visual similarity.

Qualitative analysis. Figure 3 illustrates the complementary role of intervention reconstruction beyond activation-based evidence. From the top and bottom examples alone, the feature is broadly associated with pale tail feathers, but it may also relate to tail pose, tail prominence, or in-flight status. Intervention reconstruction isolates a more specific visual change: increasing the feature brightens the tail toward white. This additional evidence produces a more precise interpretation and helps distinguish benchmark pairs whose activation patterns are visually similar.

4.2 [CLS] Intervention Evaluation

We then evaluate the effectiveness of RecSAEI on [CLS] sparse features, and more specifically, the necessity of sparse attribution and steering aggregation.



(a) (b) (c)

Figure 4: **RecSAEI provides more consistent and selective control** of [CLS] SAE features than dense steering. (a) Target [CLS] response after re-encoding the reconstructed image across intervention strengths. Dense steering leads to image collapse under a small steering strength, while RecSAEI saturates later and provides a higher peak response. (b) Target-to-off-target selectivity for different numbers of attributed patch-SAE features. The dense curve with $s > 0.01$ is dashed, as the reconstruction already collapses. (c) Target-to-off-target selectivity against the number of steered dimensions K. We select $s = \pm 0.01$ for dense and ± 0.03 for RecSAEI to roughly match the target response, as shown in Figure 4a.

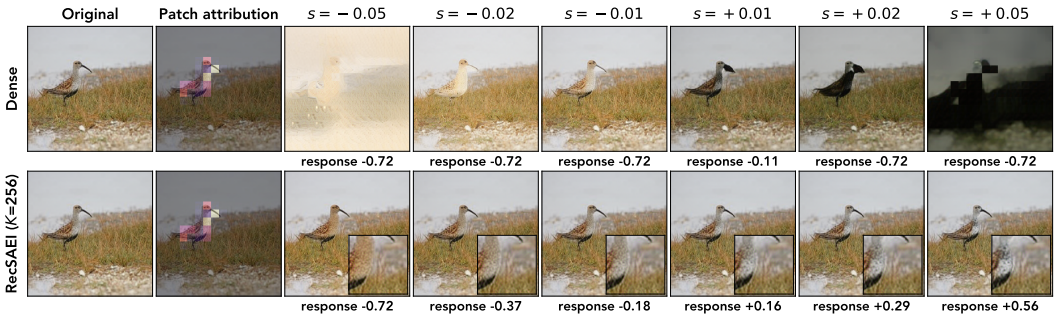


Figure 5: **Qualitative comparison of dense and sparse [CLS] steering**. Dense steering induces large global appearance changes and an inconsistent target response, whereas RecSAEI gradually changes the bird's neck color from brown to grey with black spots, with a consistent [CLS] response across intervention strengths. The original column is a direct RAE reconstruction without intervention.

Experiment setting. We evaluate [CLS] sparse feature interventions on NABirds using a frozen DINOv3 encoder and RAEv2 decoder, with separate SAEs trained on final-layer [CLS] tokens and penultimate-layer patch tokens. We sample 128 [CLS] features with five validation images per feature. For each image–feature pair, we compare the proposed sparse steering with a dense baseline that directly uses the patch-level gradient direction for steering. Both settings apply interventions on the same selected patches and share the same intervention budget, varying the number of contributing SAE features K and intervention strength. We measure round-trip activation changes on the target [CLS] dimension as well as off-target changes, using the unedited reconstruction as reference.

Results. RecSAEI provides substantially more consistent control over target [CLS] features than dense steering. As shown in Figure 4a, dense steering exhibits a narrow effective range. Increasing the intervention strength s above 0.01 leads to degraded response. In contrast, aggregating attributed patch SAE directions provides an approximately monotonic target response until $s > 0.2$ and a higher peak response. RecSAEI also improves intervention selectivity. Figure 4b shows that when more than 32 patch dimensions are aggregated, the target feature changes more strongly relative to off-target [CLS] features. Selectivity decreases slightly under larger intervention strengths, but much more slowly than dense steering. Figure 4c further shows that selectivity improves as aggregation involves more patch SAE dimensions, and 32 patch dimensions stably exceeds the corresponding dense-steering baseline. Results with more steering strengths are available in §C.4.

Qualitative analysis. Figure 5 reveals the corresponding difference in image space. Dense steering quickly introduces large global changes in illumination and appearance. The collapsed reconstruction also fails to increase the target [CLS] feature. In contrast, RecSAEI produces gradual visual changes concentrated around the attributed regions (from brown color to grey with black spots) and a consistent response in the target [CLS] activation. Figure 6 shows that increasing the number of

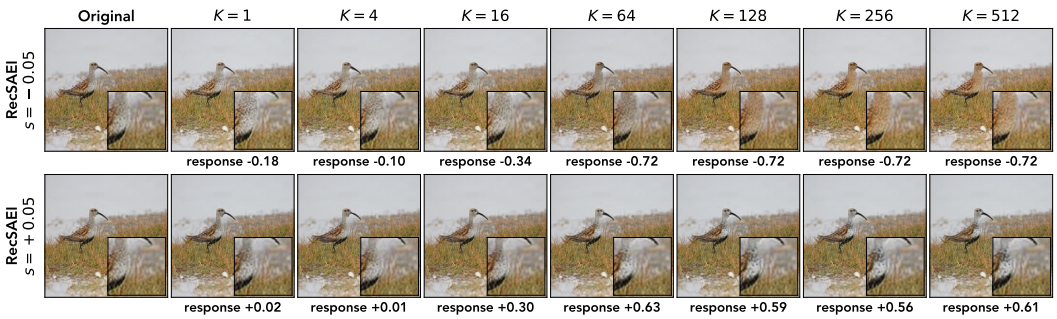


Figure 6: **Effect of the number of steered patch-SAE features** K. A small K involves a limited number of patch-SAE features and produces only slight visual changes. Increasing K makes changes more significant in both appearance and target response while preserving the overall image appearance. The intervention converges once the major contributing features are included ($K \geq 128$), validated by the target response. The original column is a direct RAE reconstruction without intervention.

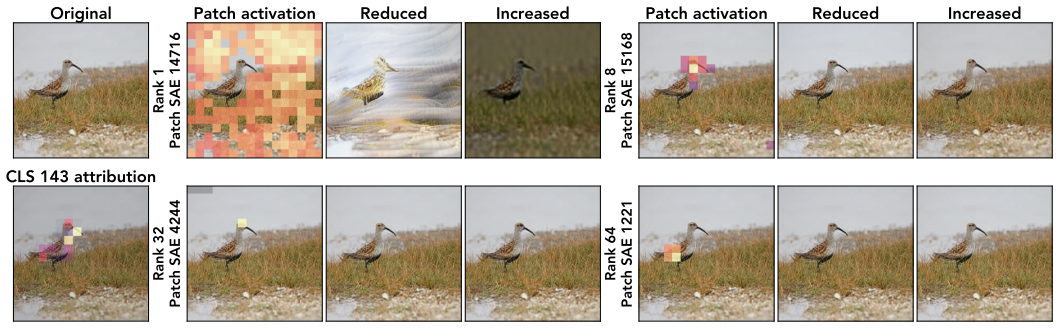


Figure 7: **Patch-SAE decomposition of a `[CLS]` SAE feature**. Highly attributed patch SAE dimensions localize to different image regions and induce distinct visual changes under intervention, revealing the local sparse components that jointly support the global `[CLS]` feature. The original column is a direct RAE reconstruction without intervention. We apply interventions to patches with the feature activated instead of attributed patches.

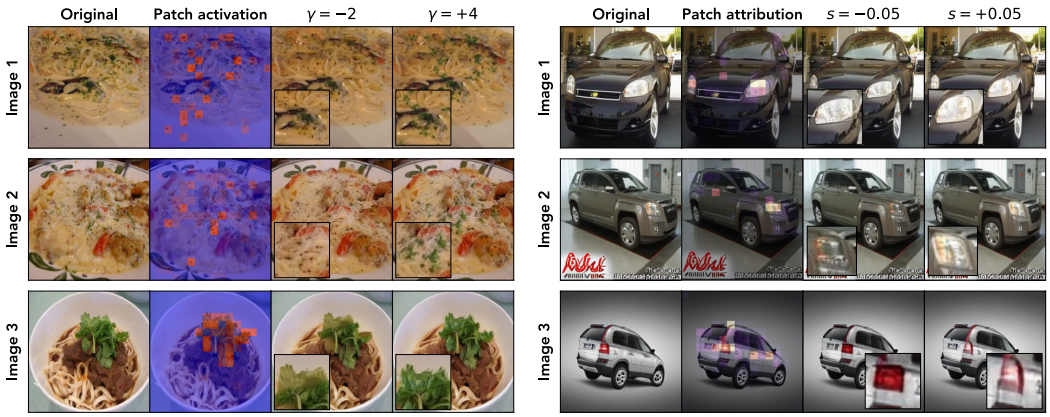
involved patch-SAE dimensions K produces more significant visual changes while preserving the overall image structure. At the same time, the induced target response change also becomes more stable. More intervention examples are in §E.

4.3 Analysis and Discussions

Decomposing `[CLS]` to Patch Sparse Features. Beyond enabling stable steering, our formulation exposes the patch-level components underlying a target `[CLS]` SAE feature. Figure 7 shows several of the highest-attributed patch SAE dimensions for the same target `[CLS]` feature. These dimensions exhibit distinct activation patterns and intervention effects, capturing heterogeneous visual properties, such as object presence, color, and shape.

The top-ranked dimension corresponds to the "elsewhere" concept, which is essential in indicating the object's presence (Fel et al., 2026). This dimension produces significant visual changes when intervened on. However, RecSAEI only modifies the attributed patch tokens (shown in Figure 5), which explains the marginal visual change in Figure 6. At the same time, this dimension also leads to relatively poor selectivity when K is small. With more dimensions involved, their aggregated direction becomes more meaningful and produces more targeted changes. We list detailed statistics of sparse dimension attribution in §D.2. The decomposition provides an interpretable account of the steering direction, connecting the response of a global `[CLS]` feature to the local visual ingredients that contribute to it. We present more attributed patch SAE dimensions and examples in §E.1.

Consistency of Intervention. Beyond producing a stable response as intervention strength varies, the attributed sparse steering direction also generalizes consistently across different images activating the same patch and `[CLS]` feature. Figure 8 shows three examples each for a shared patch and `[CLS]` SAE dimension. Although the spatial attribution varies with image content, applying the



(a) Patch SAE dimension 4443 in ImageNet–SigLIP 2. (b) [CLS] SAE dimension 412 in Cars–DINOv3.

Figure 8: **Cross-image consistency of intervention on (a) patch SAEs and (b) [CLS] SAEs**. Patch-SAE intervention uses activation to select patches for intervention, while [CLS] intervention uses gradient attribution. Both approaches identify relevant patterns, while negative and positive interventions consistently produce similar visual changes: patch dimension 4443 yields more green herbs and [CLS] dimension 412 yields taller and brighter car headlights. The original columns are direct RAE reconstructions without intervention. γ represents the scale of amplifying the target feature, and s is the ratio of the original patch embedding norm.

positive intervention generally creates more green herbs for patch SAE dimension 4443 and taller headlights for [CLS] dimension 412. The corresponding reconstructions preserve the overall scene while introducing localized appearance changes around the activated or attributed regions. This cross-image consistency suggests that the intervention captures a reusable visual direction associated with the patch or [CLS] feature.

Intervention Agreement. Beyond changes in visual appearance, we further check the agreement between the RECSAEI intervention and the response in target feature dimensions for both patch and [CLS] SAEs. As shown in Table 2, for individual patch SAE features, direction correctness and trajectory monotonicity are 64.4% and 63.1%, respectively, with a strength–response correlation of 0.71. Steering [CLS] SAE features through attributed patch features is considerably more reliable, increasing these values to 87.0%, 73.9%, and 0.93. The results suggest that, by incorporating complementary signals from patch SAE features, RECSAEI attains more stable and coherent intervention on [CLS] SAE features. At the same time, the imperfect intervention agreement also reveals the limitations of existing SAE and RAE models, motivating us to incorporate corresponding agreement signals into SAE and RAE model training, which is an important future direction.

Table 2: **Direction agreement of sparse-feature steering**. Steering generally changes the monitored feature in the intended direction, particularly for [CLS] SAE intervention, but imperfect direction agreement and trajectory monotonicity indicate that feature control is not exact.

Monitored target	Direction correctness	Trajectory monotonicity	Strength correlation
Patch SAE	64.4%	63.1%	0.71
[CLS] SAE	87.0%	73.9%	0.93

5 CONCLUSION

This paper introduces RECSAEI, an intervention-reconstruction framework for interpreting sparse visual features by decoding representation interventions into visual changes. For patch SAE features, steering along decoder directions reveals information complementary to activation-based evidence, helping disambiguate confounded features. For [CLS] SAE features, we attribute global features to patch-level dimensions and aggregate top-contributing components for steering. The sparse formulation yields stable and interpretable interventions for otherwise undecodable [CLS] SAE features. More broadly, our results motivate rethinking the format of feature explanations. We should advocate for including *how* representations change under intervention beyond static descriptions or high-activation examples to explain features comprehensively. We also need richer attribution approaches that trace how sparse features interact and compose across the network to faithfully relate visual evidence to downstream predictions.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology or experiments, implement methods, support qualitative data analysis, and interpret results. We have not used generative AI tools to generate synthetic datasets, develop conceptual models, propose or refine hypotheses, assist with translation, design methodology, or clean and reformat datasets. Developing theoretical models, formulating mathematical claims, and writing proofs are not applicable to this work. Additionally, we used generative AI tools to summarize existing literature and refine paper writing. We have reviewed all AI-assisted work, including the generated code and analysis. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work proposes to reconstruct latent-space intervention back to images. Although these reconstructions are intended for model analysis, the underlying capability to alter visual representations could potentially be repurposed for image manipulation beyond our intended use. Reconstructions reflect how the used vision encoders and RAE decoders translate the applied intervention, which should not be treated as ground-truth causal changes. Their semantics may inherit biases, artifacts, or limitations from the pretrained models and datasets. We treat reconstructions as complementary evidence for model interpretation instead of definitive descriptions. This work does not involve human-subject experiments or collection of personal data.

REPRODUCIBILITY STATEMENT

We include implementation details and benchmark construction process in §A and §B, respectively. The source code for training SAEs, applying intervention reconstruction, and pair benchmark evaluation has been attached in the supplementary material.

REFERENCES

Abubakar Abid, Mert Yuksekgonul, and James Zou. Meaningfully debugging model mistakes using conceptual counterfactual explanations. In *International Conference on Machine Learning*, pp. 66–88. PMLR, 2022.

Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11): 4311–4322, 2006.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3319–3327. IEEE, 2017.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078, 2020.

Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. In *Advances in Neural Information Processing Systems*, volume 36, pp. 25365–25389, 2023.

Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024. URL https://openreview.net/forum?id=d4dpOCqybL.

David Carlyn, Jianyang Gu, and Wei-Lun Chao. Simple localized counterfactuals for visual explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4139–4144, 2026.

Gerasimos Chatzoudis, Zhuowei Li, Gemma E Moran, Hao Wang, and Dimitris N Metaxas. Visual sparse steering: Improving zero-shot image classification with sparsity guided steering vectors. *arXiv preprint arXiv:2506.01247*, 2025.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 782–791, 2021.

Jinho Choi, Hyesu Lim, Steffen Schneider, and Jaegul Choo. Conceptscope: Characterizing dataset bias via disentangled visual concepts. In *Advances in Neural Information Processing Systems*, volume 38, pp. 9604–9639, 2026.

Arpita Chowdhury, Dipanjyoti Paul, Zheda Mai, Jianyang Gu, Ziheng Zhang, Kazi Sajeed Mehrab, Elizabeth G Campolongo, Daniel Rubenstein, Charles V Stewart, Anuj Karpatne, et al. Prompt-cam: Making vision transformers interpretable for fine-grained analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4375–4385, 2025.

Gytis Daujotas. Case study: Interpreting, manipulating, and controlling clip with sparse autoencoders. 2024a. URL https://www.lesswrong.com/posts/iYFuZo9BMvr6GgMs5/case-study-interpreting-manipulating-and-controlling-clip.

Gytis Daujotas. Interpreting and steering features in images. 2024b. URL https://www.lesswrong.com/posts/Quqekpvx8BGMMcaem/interpreting-and-steering-features-in-images.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Ieee, 2009.

Konstantin Donhauser, Kristina Ulicna, Gemma Elyse Moran, Aditya Ravuri, Kian Kenyon-Dean, Cian Eastwood, and Jason Hartford. Towards scientific discovery with dictionary learning: extracting biological concepts from microscopy foundation models. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 14321–14343, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2711–2721, 2023.

Thomas Fel, Binxu Wang, Michael Lepori, Matthew Kowal, Andrew Lee, Randall Balestriero, Sonia Joseph, Ekdeep Singh Lubana, Talia Konkle, Demba Ba, et al. Into the rabbit hull: From task-relevant concepts in dino to minkowski geometry. In *International Conference on Learning Representations*, volume 2026, pp. 154538–154584, 2026.

Javier Ferrando, Enrique Lopez-Cuena, Pablo Agustin Martin-Torres, Daniel Hinjos, Anna Arias-Duart, and Dario Garcia-Gasulla. Language models can explain visual features via steering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38947–38958, 2026.

Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3449–3457. IEEE, 2017.

Hugo Fry. Towards multimodal interpretability: Learning sparse interpretable features in vision transformers. 2024. URL https://www.lesswrong.com/posts/iYFuZo9BMvr6GgMs5/case-study-interpreting-manipulating-and-controlling-clip.

Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pp. 172–188. Springer, 2024.

Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 9277–9286, 2019.

Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*, 2019a.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pp. 2376–2384. PMLR, 2019b.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Advances in neural information processing systems*, volume 33, pp. 9841–9850, 2020.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_CDixzkzeyb.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pp. 2668–2677. PMLR, 2018.

Dahye Kim and Deepti Ghadiyaram. Concept steerers: Leveraging k-sparse autoencoders for test-time controllable generations. *arXiv preprint arXiv:2501.19066*, 2025.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pp. 5338–5348. Pmlr, 2020.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013.

Akshay Kulkarni, Tsui-Wei Weng, Vivek Narayanaswamy, Shusen Liu, Wesam A Sakla, and Kowshik Thopalli. Interpretable and steerable concept bottleneck sparse autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2918–2927, 2026.

Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. In *International Conference on Learning Representations*, volume 2025, pp. 24444–24469, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2024.

Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, July 2026. Accessed: 2026-09-19.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 95706–95742, 2025.

Vardaan Pahuja, Samuel Stevens, Alyson East, Sydne Record, and Yu Su. Automatic image-level morphological trait annotation for organismal images. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=oFRbiaib5Q.

Sangha Park, Seungryong Yoo, Jisoo Mok, and Sungroh Yoon. Save: Sparse autoencoder-driven visual information enhancement for mitigating object hallucination. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7935–7944, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PmLR, 2021.

Sukrut Rao, Sweta Mahajan, Moritz Böhle, and Bernt Schiele. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *European Conference on Computer Vision*, pp. 444–461. Springer, 2024.

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 *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Shufan Shen, Junshu Sun, Qingming Huang, and Shuhui Wang. Vl-sae: Interpreting and enhancing vision-language alignment with a unified concept set. In *Advances in Neural Information Processing Systems*, volume 38, pp. 45235–45265, 2026.

Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1532–1540. IEEE, 2021.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9240–9249. IEEE, 2020.

Sungbin Shin, Yohan Jo, Sungsoo Ahn, and Namhoon Lee. A closer look at the intervention procedure of concept bottleneck models. In *International Conference on Machine Learning*, pp. 31504–31520. PMLR, 2023.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. In *EMNLP (Findings)*, pp. 1690–1712, 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel HAZIZA, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Herve Jegou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=2NlGyqNjns. Featured Certification.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026.

David Steinmann, Wolfgang Stammer, Felix Friedrich, and Kristian Kersting. Learning to intervene on concept bottlenecks. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 46556–46571, 2024.

Samuel Stevens, Wei-Lun Chao, Tanya Berger-Wolf, and Yu Su. Interpretable and testable vision features via sparse autoencoders. *arXiv preprint arXiv:2502.06755*, 2025.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1921–1930. IEEE, 2023.

Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 595–604, 2015.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 111–119. IEEE, 2020.

Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8927–8948, 2021.

Jiaqi Xu, Cuiling Lan, and Yan Lu. Deciphering functions of neurons in vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3173–3181, 2025.

Kaichen Zhang, Yifei Shen, Bo Li, and Ziwei Liu. Large multi-modal models can interpret features in large multi-modal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3650–3661, 2025a.

Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A Ehinger, and Benjamin IP Rubinstein. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11682–11690, 2021.

Ziheng Zhang, Jianyang Gu, Arpita Chowdhury, Zheda Mai, David Carlyn, Tanya Berger-Wolf, Yu Su, and Wei-Lun Chao. Finer-cam: Spotting the difference reveals finer details for visual explanation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9611–9620, 2025b.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, volume 2026, pp. 35791–35820, 2026.

Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Revisiting the importance of individual units in cnns via ablation. *arXiv preprint arXiv:1806.02891*, 2018.

Table of Contents in Appendix

A Implementation Details **16**

A.1 SAE Training 16

A.2 SAE Statistics 16

A.3 Steering Implementation 16

A.4 Metric Definition 17

B Paired Benchmark Details **18**

B.1 Benchmark Constraints 18

B.2 Evaluation Implementation 18

B.3 Prompts for Different Image Input 18

B.4 Text Description Generation Prompt 19

B.5 Prediction Prompt 19

C Ablation Study **20**

C.1 Activation Indication Form 20

C.2 Comparisons Using CLIP and IoU Scores 20

C.3 Comparison Between Text and Image Explanation 20

C.4 Selectivity on More Steering Strengths 21

C.5 Reconstruction Using 128 Patch SAE Dimensions 21

C.6 Attribution Faithfulness 22

C.7 Comparison between Attribution Methods 22

C.8 Comparison with Contrastive Steering 23

D More Quantitative Results **24**

D.1 Trace the Steering Changes 24

D.2 `[CLS]` Attribution Statistics 24

E More Qualitative Results **25**

E.1 `[CLS]` Dimension Decomposition 25

E.2 More Examples of Intervention Reconstruction 25

E.3 More Consistency Examples for `[CLS]` SAE Intervention 25

A Implementation Details

A.1 SAE Training

We train SAE models for two architectures: DINOv3-L/16 (Siméoni et al., 2026) and SigLIP2-L/16 (Tschannen et al., 2025). Both of them use a ViT-large backbone and a patch size of 16. Because SigLIP 2 does not have a conventional `[CLS]` token, we only train `[CLS]` SAEs for DINOv3.

Patch SAE. We train the patch SAE model on the penultimate-layer (layer -2) activations extracted from images resized to 256×256. It operates on the 256 spatial patch tokens per image, using a 1,024-dimensional input and a 16,384-dimensional dictionary. We used a BatchTopK SAE (Bussmann et al., 2024) with $k = 32$, Adam optimization, and a learning rate of 4×10^{-4}. The training budget was 100 million token presentations for NABirds and ImageNet, and 25 million for CUB and Cars. Before training, we shuffle the patch activations to mitigate location-induced correlations, followed by ordered inference on the training and validation sets.

`[CLS]` SAE. We train the `[CLS]` SAE separately on the final-block (layer -1) `[CLS]` token activations. It uses a 1,024-dimensional input, a 4,096-dimensional dictionary, and the same BatchTopK objective with $k = 32$, together with Adam and a learning rate of 4×10^{-4}. Because each image only contributes one `[CLS]` token, training uses a batch size of 512 and two million token presentations. We set the dead-feature threshold to 100,000 presentations. We keep both SAEs frozen during the subsequent attribution and steering experiments.

A.2 SAE Statistics

Table 3: **Validation statistics of patch SAEs across datasets and visual backbones**. We report reconstruction MSE, L_1 statistics, and the numbers of dead, almost-dead, and dense features for SAEs trained on DINOv3 and SigLIP 2 representations.

Model	Dataset	MSE	L_1	Dead	Almost dead	Dense
DINOv3	NABirds	62.15	2,579.18	0	0	432
	CUB	72.78	2,517.16	1	1	473
	Cars	75.49	2,577.52	2	2	363
	ImageNet	114.62	2,813.61	55	166	407
SigLIP 2	NABirds	0.26	406.78	105	105	475
	CUB	0.31	358.01	78	78	544
	Cars	0.31	298.32	85	85	484
	ImageNet	0.35	257.74	231	600	431

Table 3 summarizes the validation statistics of the patch SAEs across datasets and backbones. We use activation frequency below 10^{-7} as the criterion for almost-dead dimensions and above 10^{-2} for dense dimensions. The number of dense features remains relatively stable across datasets for both backbones. Relatively, ImageNet has more tokens and is thus less optimized. These results indicate that the learned sparse representations are broadly well utilized.

A.3 Steering Implementation

We use dinov3l-k1 and siglip2l-k1 RAEv2 decoders for DINOv3 and SigLIP 2, respectively.

Patch SAE. For interventions on a single patch-SAE dimension, each image is first encoded to the SAE input layer and transformed into patch-level SAE codes $\boldsymbol{z}_p$. Given a target dimension k, we modify only that coordinate by setting $\boldsymbol{z}'_{p,k} = \gamma \boldsymbol{z}_{p,k}$, while leaving all other SAE dimensions unchanged. We then decode both the original and edited SAE codes and add the decoder difference, $D_{\text{SAE}}(\boldsymbol{z}'_p) - D_{\text{SAE}}(\boldsymbol{z}_p)$, to the original dense patch representations, thereby avoiding errors from SAE reconstructions. The intervention is restricted to the spatial patches containing 80% of the dimension's positive activation mass, with at least four patches selected. We leave the `[CLS]` and register tokens unmodified. We pass the patch embeddings through the remaining encoder and use the RAE decoder to map the intervened embeddings back to an image. For the direction agreement

experiment in Table 2, we re-encode the produced images through the original encoder and patch SAE. We evaluate suppression and amplification using scales $\gamma \in \{0, 0.5, 1, 2, 4\}$, where $\gamma = 0$ removes the feature and $\gamma > 1$ amplifies it.

[CLS] SAE. We do not directly edit [CLS] SAE dimensions, as the RAE decoder does not take [CLS] embeddings as inputs. Instead, we treat each target [CLS] SAE dimension as a scalar objective and intervene upstream on the penultimate-layer patch representations that influence it. For each target dimension, we compute its gradient with respect to the raw patch tokens $\boldsymbol{g}_p = \nabla_{\boldsymbol{h}_p} y_c$. As in patch-SAE intervention, we steer only the smallest set of patches containing 80% of the gradient mass. The sparse attribution first decomposes gradients into patch-SAE dimensions by decoder-direction alignment $q_{p,k} = \boldsymbol{g}_p^\top \boldsymbol{w}_k$, which indicates their sensitivity to the target change. We rank patch-SAE dimensions by the product of dimension activation and sensitivity and select the top K dimensions for the subsequent intervention. The aggregation step combines their decoder directions using the signed gradient weights and normalizes the directions across patches. Finally, the intervention acts similarly to the patch-SAE intervention and applies a steering strength toward the aggregated direction to the selected patch embeddings. Here, for better control, we define the intervention magnitude relative to the Frobenius norm of the original patch-token tensor. This way, dense steering and SAE-derived steering receive the same amount of input-level changes.

We propagate the updated patch tokens through the remaining encoder and RAE decoder, keeping [CLS] and register tokens unchanged. We then re-encode the edited image and measure changes in the target [CLS] SAE feature's dense and sparse activations h and f, together with off-target changes and selectivity. The intervention controls a [CLS] SAE feature indirectly through an attributed subspace of patch-SAE decoder directions.

A.4 Metric Definition

Normalized target response. We measure the relative changes in the target [CLS] dimension: $R_{\text{target}} = \Delta \boldsymbol{z}_{\text{target}} / \boldsymbol{z}_{\text{target}}$, where $\Delta \boldsymbol{z}_{\text{target}}$ is the absolute change after the edited image is decoded and re-encoded, relative to the re-encoded no-op reconstruction, and $\boldsymbol{z}_{\text{target}}$ is the target's original activation on the input image. We report the round-trip response in the main paper. For signed interventions, the intended response is $\text{sign}(s) R_{\text{target}}$, where s is the steering strength, so positive values indicate movement in the requested direction.

Target selectivity. We measure the relative changes in the target dimension versus other dimensions. Specifically, we divide the absolute target change by the mean absolute change of the other [CLS] SAE dimensions. Values above one indicate that the target changes more than an average off-target dimension. Larger values indicate better selectivity.

Directional correctness. We apply both amplified and suppressed interventions on the SAE activations and measure whether the resulting change in the target dimension matches the intended direction. We report the fraction of nonzero-strength interventions for which $s \Delta \boldsymbol{z}_{\text{target}} > 0$.

Monotonicity. Beyond directional correctness, we measure stricter monotonicity, where increasing the signed intervention strength consistently increases the target SAE response. Specifically, we ask that every adjacent response change increases by more than 10^{-6}. We calculate the metric on the round-trip responses for both patch and [CLS] SAEs and report the fraction of target trajectories with different steering strengths whose response is nondecreasing.

Strength correlation. We report the mean Spearman rank correlation between the signed intervention strength and the resulting target response. For each trajectory from one SAE dimension and one image, we collect the requested strengths s and the corresponding round-trip target response $\Delta \boldsymbol{z}(s)$. We then compute

$$\rho_{\text{trajectory}} = \text{SpearmanCorr}\big([s_1, \ldots, s_n], [\Delta \boldsymbol{z}(s_1), \ldots, \Delta \boldsymbol{z}(s_n)]\big).$$

Thus, negative strengths should produce negative target responses and positive strengths positive responses. A value near 1 indicates a monotonic, correctly oriented response curve; 0 indicates weak or non-monotonic dependence; and negative values indicate an inverted relationship. The correlation is computed on ranks, so it measures ordering rather than assuming a linear response and is therefore less strict than monotonicity.

B PAIRED BENCHMARK DETAILS

This section describes how we build the paired benchmark to evaluate the intervention on single patch-SAE dimensions.

B.1 BENCHMARK CONSTRAINTS

Table 4: The thresholds for the paired benchmark construction.

Constraint	**DINOv3**	**SigLIP 2**
Max low/high activation ratio	0.15	0.15
Min high-image percentile	0.85	0.85
Min activation-percentile gap	0.60	0.50
Min training activation	1.0	0.1
Dense-similarity baseline percentile	$99th$	$99th$

Our overall target for the selected pairs is that they are *similar in their overall representation but differ clearly along the target SAE feature*. First, we require each pair to have dense-embedding similarity above the $99th$ percentile of randomly sampled image pairs, which reduces the risk that the pair can be easily distinguished by coarse differences unrelated to the target feature. For SAE activation distinction, we require the high-activation image to lie above the 85th activation percentile, ensuring sufficient signals. The low-to-high activation ratio must be below 0.15, enforcing a strong relative contrast within the pair. Complementarily, the activation-percentile gap must exceed 0.60 for DINOv3 and 0.50 for SigLIP 2, ensuring that the two images are also well separated relative to the feature's overall activation distribution. Finally, we require the high image to reach a minimum training activation of 1.0 for DINOv3 and 0.1 for SigLIP 2 to exclude dimensions whose positive examples remain weakly expressed. We select 5 visually matched pairs for each tested SAE dimension.

B.2 EVALUATION IMPLEMENTATION

After acquiring the SAE trained on the target dataset, we first filter out dense, dead, and almost dead dimensions, as well as those with stronger activation on backgrounds. We then randomly select $1{,}000$ dimensions from the remaining pool that can produce 5 image pairs meeting the requirements in §B.1. Together, we evaluate $5{,}000$ image pairs for each dataset–encoder combination. Given a specific type of feature explanation, the benchmark asks GPT-5.6-Terra to predict the image with higher activations on the target feature. Explanations are presented in two formats: images and text descriptions.

We test four different groups of image explanations: (1) images with top-ranked activations and the corresponding overlay image showing the patches with high activations, denoted as "Top"; (2) Top examples and the corresponding reconstructions with the target feature amplified and suppressed; (3) Top examples and bottom examples, *i.e.*, images with the lowest activation in the target feature; (4) the combination of top, bottom examples, and the intervention reconstruction on top examples. When used as the input, we adopt the prompts presented in §B.3.

In Table 1, we only present the text description generated based on top examples. Table 7 evaluates text descriptions generated based on different image inputs. We use the image input prompts and the text description generation instruction in §B.4 to produce descriptions.

Given the explanation and test image pair, we use the prompt in §B.5 to produce the prediction and calculate the accuracy in predicting the image with higher activation.

B.3 PROMPTS FOR DIFFERENT IMAGE INPUT

Top input instruction

The following examples provide a feature explanation. Each example contains an original image and its activation overlay, where redder regions have higher activation. Infer the visual property consistently highlighted by the overlays.

Top + intervention input instruction

The following examples provide a feature explanation. Each example contains the original image, its activation overlay (redder regions have higher activation), a reconstruction with activation reduced, and a reconstruction with activation increased. Infer the visual property indicated by both the overlay and the consistent change from the reduced to the increased reconstruction.

Top + bottom input instruction

The following examples provide a feature explanation. Each reference contains a high-activating top image and a low-activating bottom image. Infer the visual property that consistently distinguishes the top examples from the bottom examples.

Top + bottom + intervention input instruction

The following examples provide a feature explanation. Each reference contains a high-activating top image, its activation overlay (redder regions have higher activation), a low-activating bottom image, a reconstruction from a reduced intervention, and a reconstruction from an increased intervention. Infer the visual property that distinguishes the top examples from the bottom examples and is supported by both the overlays and the consistent reconstruction changes.

B.4 Text Description Generation Prompt

Text description generation instruction

Return exactly one standalone, concise sentence describing the inferred feature. Be as visually specific as the evidence permits, including concrete color, texture, shape, location, body part, or semantic properties when supported. Do not mention the references, examples, heatmaps, activations, or inference process.

B.5 Prediction Prompt

Prediction instruction

Given this feature explanation, which image has a higher activation for that feature? Return exactly one letter: A or B.

C Ablation Study

C.1 Activation Indication Form

Table 5: Comparison between different forms of indicating the activation region in the top examples. This experiment uses 200 individual SAE features (1,000 image pairs) extracted from NABirds using DINOv3 as the encoder.

Activation	Explanation Method	
	Top	Top + Intervention
Contour	67.0	70.5
Overlay	70.6	72.1

In the main experiments, we use an overlay to indicate the patches within an image that have high activation in the target sparse feature, with red color for higher activation and blue color for lower activation. We compare this design choice with using colored contours around the high-activating patches and report the results in Table 5. For both top and top + intervention explanations, the overlay yields a consistent performance advantage, which may be because the overlay provides clearer differentiation between high- and low-activating regions, especially when there is a larger ratio of patch tokens with high activation.

C.2 Comparisons Using CLIP and IoU Scores

Table 6: **Standard similarity metrics do not distinguish explanation quality**. CLIP and IoU scores remain nearly unchanged across explanation settings, despite adding bottom examples and intervention reconstructions that provide qualitatively different semantic evidence. This experiment uses 200 dimensions (1,000 image pairs).

Explanation approach			NABirds-DINOv3	
Top	Bottom	Intervention	CLIP Score	IoU Score
✓			0.204	0.240
✓		✓	0.202	0.234
✓	✓		0.202	0.248
✓	✓	✓	0.202	0.232

Table 7: **Comparison between using text and image as the explanation format**. Directly using example images consistently outperforms the text descriptions derived from the same examples. The gap widens as accuracy increases. This experiment uses 200 dimensions (1,000 image pairs) from NABirds–DINOv3.

Explanation approach			NABirds-DINOv3		
Top	Bottom	Intervention	Text	Image	Δ
✓			62.9	**63.4**	↑ 0.5
✓		✓	64.7	**65.4**	↑ 0.7
✓	✓		67.3	**69.4**	↑ 2.1
✓	✓	✓	67.1	**72.9**	↑ 5.8

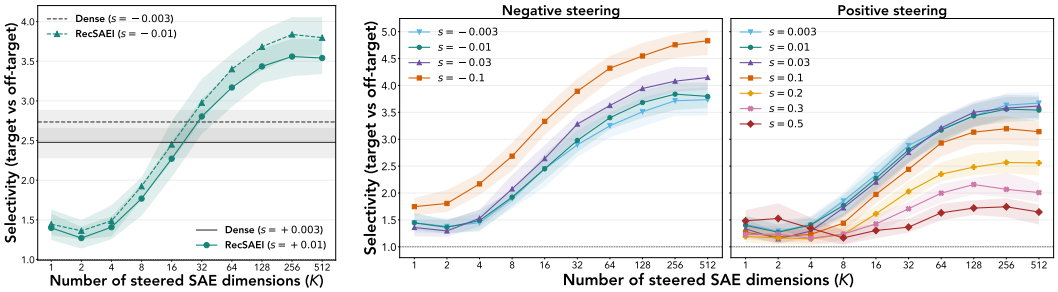
Prior work employs CLIP and IoU scores to evaluate whether the produced text descriptions can faithfully match the corresponding image regions with high activation (Xu et al., 2025; Zhang et al., 2025a). CLIP and GroundingDINO models provide accurate recognition and localization for general objects, such as ImageNet categories. However, they may be unreliable when dealing with fine-grained attributes, which are out of their training distribution. We generate the text descriptions given different image inputs using the prompts in §B.4 and report their resulting CLIP and IoU scores in Table 6. The experiment runs on NABirds. Across the four explanation settings, the CLIP score varies only from 0.202 to 0.204, while IoU ranges from 0.232 to 0.248, with no consistent improvement as additional explanatory evidence is introduced. These results contrast with our main results in Table 1, where different explanation methods yield large performance differences in identifying feature semantics. This suggests that generic text–image similarity and overlap-based metrics primarily capture coarse semantic correspondence and are insufficiently sensitive to the fine-grained distinctions needed to evaluate sparse-feature explanations.

C.3 Comparison Between Text and Image Explanation

We further compare text descriptions generated based on different image explanation inputs in Table 7. Intervention reconstruction improves the prediction accuracy when only top examples are used to derive the feature description. However, the description does not benefit further from intervention

when bottom examples are also provided. It can be understood that a concise sentence describing the inferred feature has limited information capacity. In comparison, image explanations, when presented directly to the evaluator, can create more comprehensive feature correspondence with the test image pairs. Notably, as more image explanations are provided, the performance gap between using images directly and using text descriptions also grows. The results support that images provide a more flexible and comprehensive reasoning space compared with text descriptions.

C.4 Selectivity on More Steering Strengths



(a) (b)

Figure 9: **Effect of steering strength on intervention selectivity**. (a) Selectivity comparison using steering strengths ($|s| = 0.003$ for dense steering and $|s| = 0.01$ for RecSAEI) with roughly matched target response. RecSAEI remains more selective across different numbers of source dimensions K. (b) Selectivity of RecSAEI across intervention strengths. For negative steering, stronger interventions consistently increase selectivity, whereas positive steering achieves its highest selectivity at relatively small strengths and becomes less selective as the intervention grows stronger. Shaded regions indicate variability across evaluated examples.

We further study how steering strength affects the selectivity of SAE intervention. Figure 9a provides an additional comparison with dense steering. In Figure 4c, we compare dense and RecSAEI using fixed strengths of 0.01 and 0.03, respectively. Because the two steering spaces can produce different response magnitudes at the same nominal scale, here we choose another group of strengths that yield approximately matched target responses ($|s| = 0.003$ for dense steering and $|s| = 0.01$ for RecSAEI). Under this response-matched setting, RecSAEI still achieves higher selectivity as K increases over 32, indicating a consistent advantage of the proposed RecSAEI.

Figure 9b varies the intervention strength s while increasing the number of attributed patch SAE dimensions K. For negative steering, stronger interventions consistently produce greater separation between the target `[CLS]` feature and off-target features. This trend remains stable across K, with selectivity generally increasing as more attributed source dimensions are incorporated. For positive steering, relatively small strengths, such as $s = 0.003$, 0.01, and 0.03, produce similar and consistently high selectivity as K increases. In contrast, further increasing the strength progressively reduces selectivity: although the target feature continues to be promoted, the intervention increasingly affects off-target features as well. Overly strong interventions move the representation beyond the training distribution and introduce broader visual changes.

C.5 Reconstruction Using 128 Patch SAE Dimensions

We qualitatively compare dense steering and SAE steering using 256 patch-SAE features in Figure 5. Here, we also show the comparison using 128 patch-SAE features in Figure 10. We derive similar conclusions as in the main paper: Dense patch steering is markedly more sensitive to intervention strength than the activation-weighted SAE direction. Even at moderate strengths, the visual changes extend beyond the attributed appearance, while stronger interventions rapidly introduce severe color shifts and reconstruction artifacts. At the same time, the target response decreases quickly due to collapsed visual reconstruction. In contrast, RecSAEI produces a substantially smoother progression across s, with visual changes remaining concentrated on the relevant appearance and the target response varying more consistently.

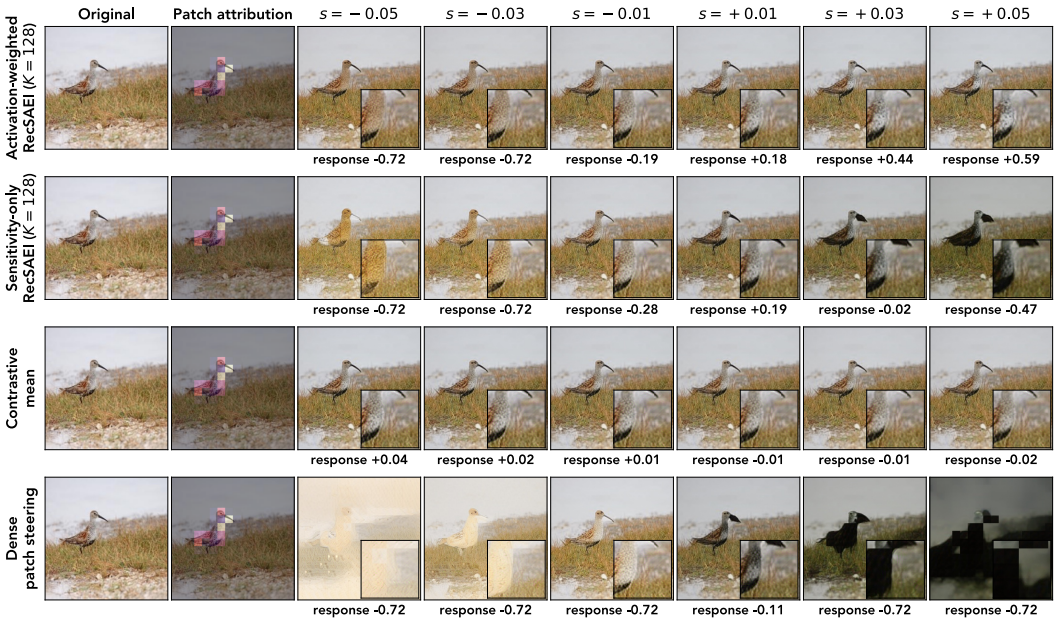


Figure 10: **Comparison of intervention reconstructions across intervention strengths**. We vary the steering strength s from negative to positive while keeping the intervention region fixed by the patch attribution. The first row uses our activation-weighted SAE steering with $K = 128$ source features. The remaining rows replace the steering direction with sensitivity-only SAE aggregation, a contrastive-mean direction, and dense patch steering, respectively. The value below each reconstruction reports the target `[CLS]` feature response.

C.6 Attribution Faithfulness

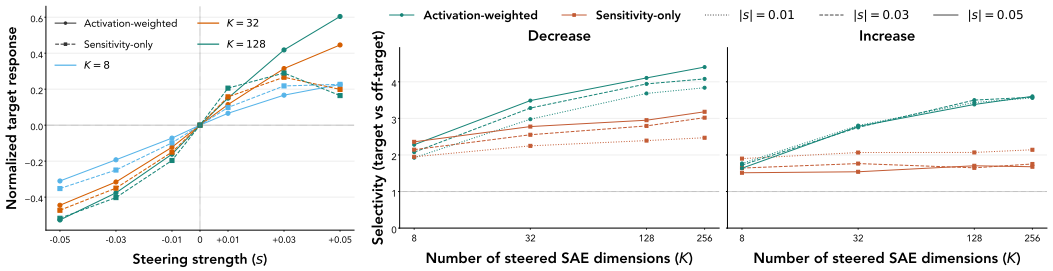
Table 8: **Target `[CLS]` response after steering different patch SAE dimensions**. Steering top-attributed dimensions produces substantially larger `[CLS]` activation changes than steering random active or bottom-attributed dimensions.

Source dimensions	Mean absolute `[CLS]` change	Mean token intervention	Change ratio
Top attributed	7.70	396.4	1.94%
Random active	0.23	92.8	0.20%
Bottom attributed	0.01	35.3	0.02%

To test whether our attribution scores identify patch SAE features that functionally support the target `[CLS]` feature, we steer dimensions selected by different criteria and measure the resulting target response. As shown in Table 8, top-attributed dimensions produce a much larger mean absolute `[CLS]` change than randomly selected active dimensions and bottom-attributed dimensions. While these dimensions usually have larger activations and contribute larger interventions, their change ratio per absolute intervention is also larger. This large separation indicates that the proposed attribution identifies sparse patch features whose intervention has a substantially stronger effect on the target `[CLS]` feature. We ablate the attribution approach in §C.7.

C.7 Comparison between Attribution Methods

Our `[CLS]` intervention constructs its steering direction from patch SAE features using both their sensitivity to the target `[CLS]` feature and their activation in the current image. Sensitivity captures how strongly changing a patch feature could affect the target, but a highly sensitive feature need not be active in the selected image region. Steering such features may therefore inject directions that are relevant to the target in principle but are not grounded in the image's current sparse representation.



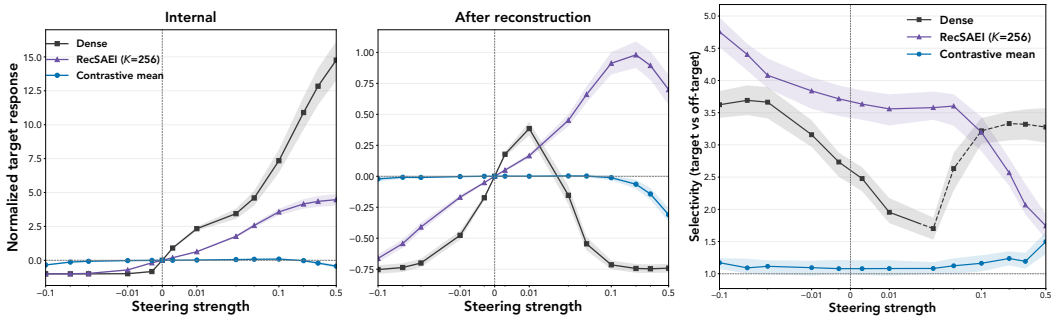
(a) (b)

Figure 11: **Comparison between the proposed activation-weighted dimension attribution and sensitivity-only attribution**. (a) Normalized target-feature response under different steering strengths. Sensitivity-only attribution yields larger responses at small strengths but also saturates earlier than activation-weighted attribution. (b) Selectivity for decreasing and increasing interventions. Across steering magnitudes and number of used patch-SAE features, activation-weighted attribution achieves consistently higher selectivity than sensitivity-only steering, with the gap widening as K grows.

We compare our activation-weighted attribution against a sensitivity-only variant in Figure 11a and Figure 11b. When only the most sensitive patch-SAE features are used, sensitivity-only attribution produces slightly larger target response and higher selectivity. However, as the steering strength increases, sensitivity-only attribution saturates more quickly. As shown in the second row of Figure 10, sensitivity-only steering produces a similar visual trend at small intervention strengths but becomes substantially less stable as the magnitude increases.

The selection diagnostics provide a direct explanation for this behavior. At $K = 128$, 96.0% of the dimensions selected by sensitivity-only steering are inactive throughout the selected patch mask. Sensitivity-only steering therefore attempts to manipulate features that are largely absent from the sparse representation of the image. As a result, it edits directions that may be locally sensitive in principle but are not grounded in the image's actual visual evidence, which weakens both controllability and selectivity. In contrast, activation-weighted steering favors source dimensions that are both attributed to the target and active in the selected region, producing a steering direction that is better aligned with the image content and therefore yields stronger, more selective interventions.

C.8 Comparison with Contrastive Steering



(a) (b)

Figure 12: **Comparison between RecSAEI and contrastive steering** derived from the mean feature difference between samples with high and low activation. (a) Normalized target response, measured internally before reconstruction and after reconstruction. Dense steering produces the largest internal response, but much of this effect does not survive reconstruction. In contrast, RecSAEI yields a smaller internal response but a substantially stronger and more consistent post-reconstruction target response. The contrastive-mean direction produces almost no controllable response in either space. (b) Selectivity of the reconstructed interventions. RecSAEI is consistently more selective than both dense steering and the contrastive-mean approach.

Beyond deriving steering directions from gradients, the difference between samples with high and low activation is another option (named "Contrastive mean"). We compare the intervention effects both before and after reconstruction for dense, SAE, and contrastive mean in Figure 12. In the internal representation, dense steering produces the largest target response. However, after reconstruction, the picture changes substantially, leading to decreased target response. By contrast, RecSAEI produces a much stronger post-reconstruction response over a wide range of positive strengths. Thus, RecSAEI transfers more effectively into the reconstructed image space, and contrastive mean only produces marginal changes in the target response, which is confirmed in Figure 10. Additionally, contrastive-mean stays close to a selectivity of 1, indicating that it changes the target little more than off-target features and therefore does not isolate the feature of interest.

D More Quantitative Results

D.1 Trace the Steering Changes

Table 9: **Round-trip stability of dense steering and RecSAEI under matched injected perturbation magnitude**. We report the resulting patch-feature change after propagation through the remaining encoder block, the target `[CLS]`-feature change before reconstruction, the target change after reconstruction and re-encoding, and the image-space MSE relative to direct reconstruction.

Steering Strength	Method	Injected patch L_2	Final patch L_2	Internal `[CLS]` target change	Round-trip `[CLS]` target change	Image MSE
$s = 0.01$	Dense	140.8	4,163.9	180.9	28.7	4.6×10^{-3}
	RecSAEI $K = 8$	140.8	419.3	21.1	4.7	2.3×10^{-5}
	RecSAEI $K = 128$	140.8	418.7	44.8	11.4	3.3×10^{-5}
	RecSAEI $K = 256$	140.8	441.5	50.7	12.6	3.9×10^{-5}
$s = 0.03$	Dense	422.4	16,024.6	342.5	-18.5	5.5×10^{-2}
	RecSAEI $K = 8$	422.4	1,448.0	30.3	13.7	2.0×10^{-4}
	RecSAEI $K = 128$	422.4	1,456.8	91.1	31.9	2.8×10^{-4}
	RecSAEI $K = 256$	422.4	1,575.9	99.8	33.9	3.3×10^{-4}

Beyond the response in the target `[CLS]` feature, we further investigate how the applied intervention propagates through the pipeline. As shown in Table 9, we start from the original injected patch embedding intervention. The final patch L_2 measures the change in patch embeddings after the final transformer layer. The final patch embeddings also serve as the input to the RAE decoder. For the target `[CLS]` feature, we measure the change after the final transformer layer (Internal) and the activation after RAE reconstruction and re-encoding using the same encoder.

Despite receiving the same initial perturbation budget, dense steering is strongly amplified by the remaining encoder computation: its final patch L_2 is roughly an order of magnitude larger than that of RecSAEI, and it produces substantially larger internal `[CLS]` responses. These large internal changes, however, do not translate faithfully through reconstruction. At $s = 0.01$, the dense target response drops from 180.9 internally to 28.7 after the round trip, while also producing an image MSE over two orders of magnitude larger than RecSAEI. The instability becomes more pronounced at $s = 0.03$: although dense steering induces an internal target change of 342.5, the round-trip response reverses sign to -18.5, accompanied by a large image deviation. In contrast, RecSAEI produces smaller but more stable perturbations: increasing K strengthens both the internal and round-trip target responses while keeping reconstruction error low. These results suggest that SAE decoder directions produce more stable intervention effects.

D.2 `[CLS]` Attribution Statistics

The experiments in §C.7 compare the proposed activation-weighted attribution with using sensitivity only. We further list the detailed statistics of sensitivity and activation values for the used patch-SAE features in Table 10. The sensitivity $q_{p,k}$ measures the alignment between patch-level gradient $\boldsymbol{g}_p$ and the decoder direction $\boldsymbol{w}_k$. As gradient magnitude is not excluded from the calculation, the sensitivity scale is relatively small. Overall, attribution rank and sensitivity rank differ substantially. Dimension 14716 is ranked top under both measures. Visualizations indicate that the feature represents the

Table 10: **Top attributed patch SAE features** for [CLS] 143 intervention. We report the top 8 patch-SAE dimensions ranked by the attribution score used in our activation-weighted steering direction, together with their average activation within the selected patch mask, average absolute sensitivity to the target [CLS] feature, and their rank under sensitivity-only selection.

Rank	Patch-SAE dimension	Attribution score	Average activation	Average absolute sensitivity	Sensitivity Rank
1	14716	79.14	339.67	0.016	3
2	10967	24.86	238.85	0.004	146
3	9392	18.21	59.24	0.008	20
4	7323	16.43	33.04	0.006	51
5	4630	10.13	77.92	0.005	71
6	12212	7.02	19.03	0.007	37
7	5179	6.06	11.48	0.005	99
8	15168	5.91	56.13	0.003	203

"elsewhere" concept (Fel et al., 2026). The intervention and this attribution analysis further support that the "elsewhere" concept is essential in representing the presence of an object.

E More Qualitative Results

E.1 [CLS] Dimension Decomposition

We present a more comprehensive decomposition result for [CLS] dimension 143 in Figure 13 and two other [CLS] features in Figure 14 and Figure 15. Across these examples, the highest-ranked patch-SAE feature is consistently dimension 14716, the "elsewhere" concept. The activation overlay shows the normalized activation. Therefore, while the figure suggests the absence of the feature on the object, the activation is merely weaker than background patches. Independently intervening on this feature alters the appearance of nearly the entire image. Activation-weighted attribution identifies the feature as relevant where it overlaps the target [CLS] attribution, but the resulting decoder contribution is applied only within the derived patch mask. We also incorporate more patch-SAE features as K increases, helping attain more stable intervention effects. Nevertheless, the decomposition provides meaningful information on how patch-level features contribute to higher-level [CLS] semantics.

E.2 More Examples of Intervention Reconstruction

We present more examples of intervention reconstruction on patch-SAE features from Figure 16 to Figure 23, and on [CLS] SAE features from Figure 24 to Figure 27.

E.3 More Consistency Examples for [CLS] SAE Intervention

We present more examples of intervening on [CLS] SAE dimensions for multiple images in Figure 28. The color of the attributed patches becomes lighter as the steering strength increases for dimension 2434. In comparison, dimension 3348 focuses on a white strip on the duck's face. Amplifying the feature will make the strip more prominent. These examples prove that the intervention leads to consistent changes in the target [CLS] feature.

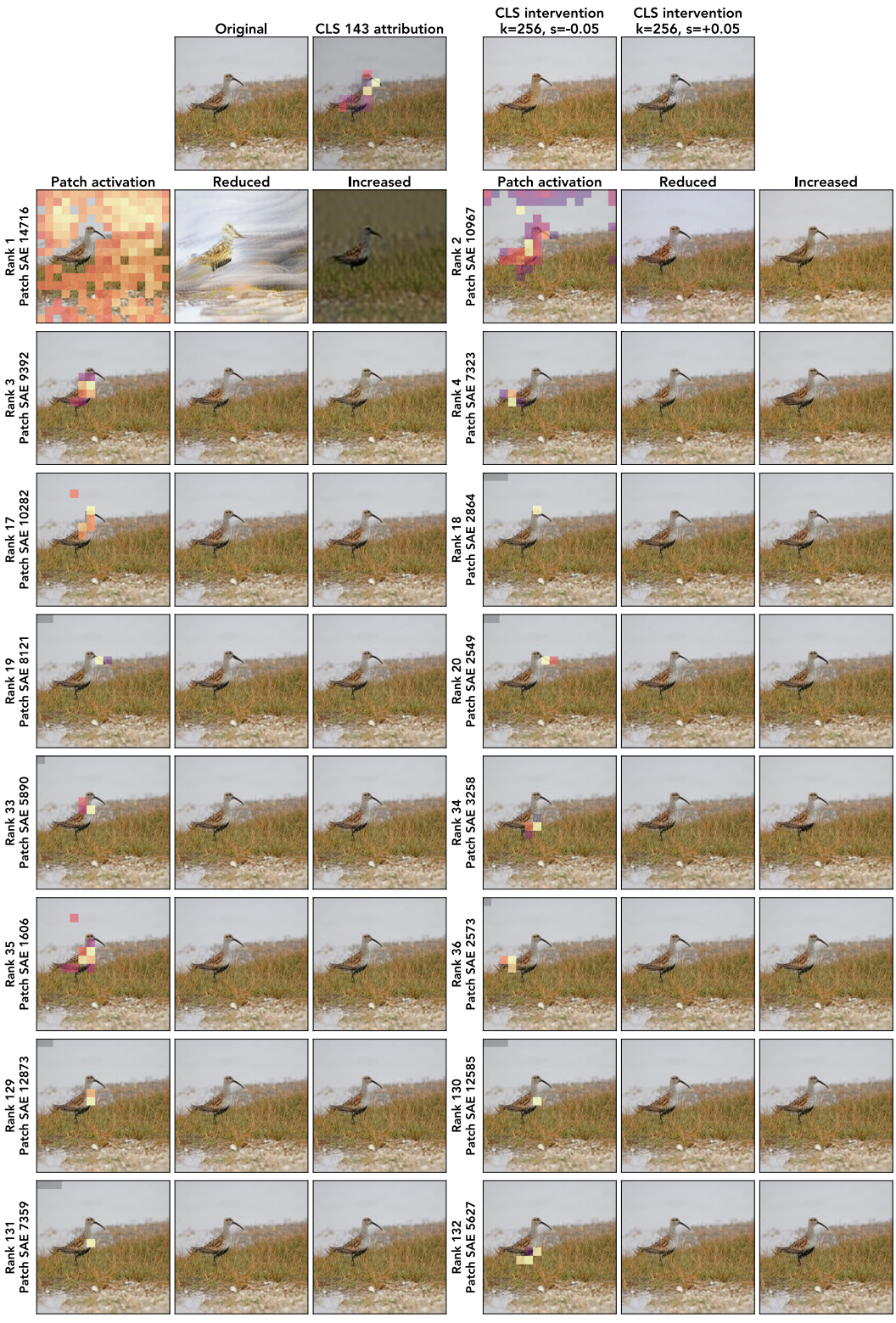


Figure 13: **Attributed Patch-SAE decomposition of `[CLS]` SAE feature** 143. Highly attributed patch-SAE dimensions localize to different image regions and induce distinct visual changes under intervention, revealing the local sparse components that jointly support the global `[CLS]` feature.

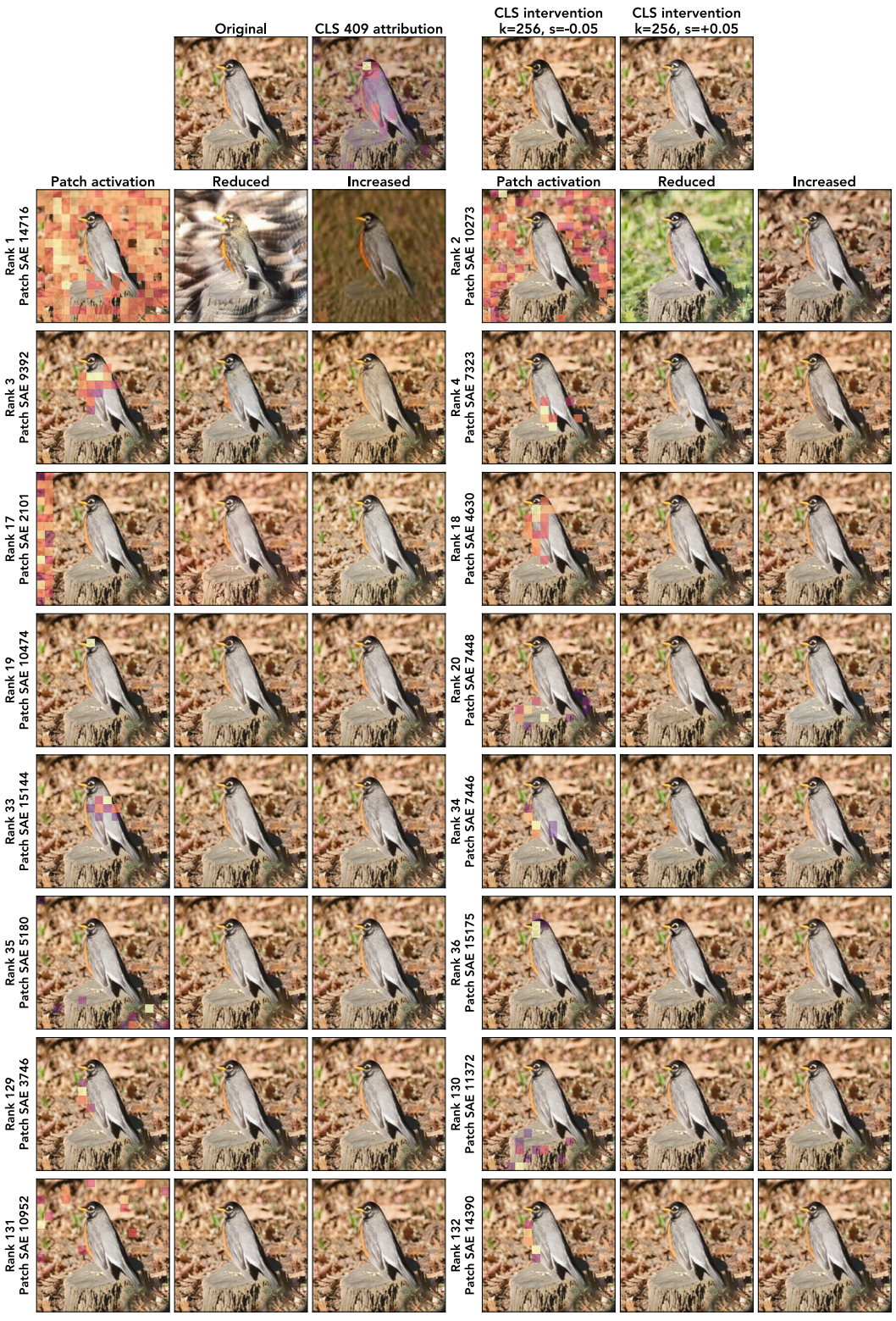


Figure 14: **Attributed Patch-SAE decomposition of `[CLS]` SAE feature** 409. Highly attributed patch SAE dimensions localize to different image regions and induce distinct visual changes under intervention, revealing the local sparse components that jointly support the global `[CLS]` feature.

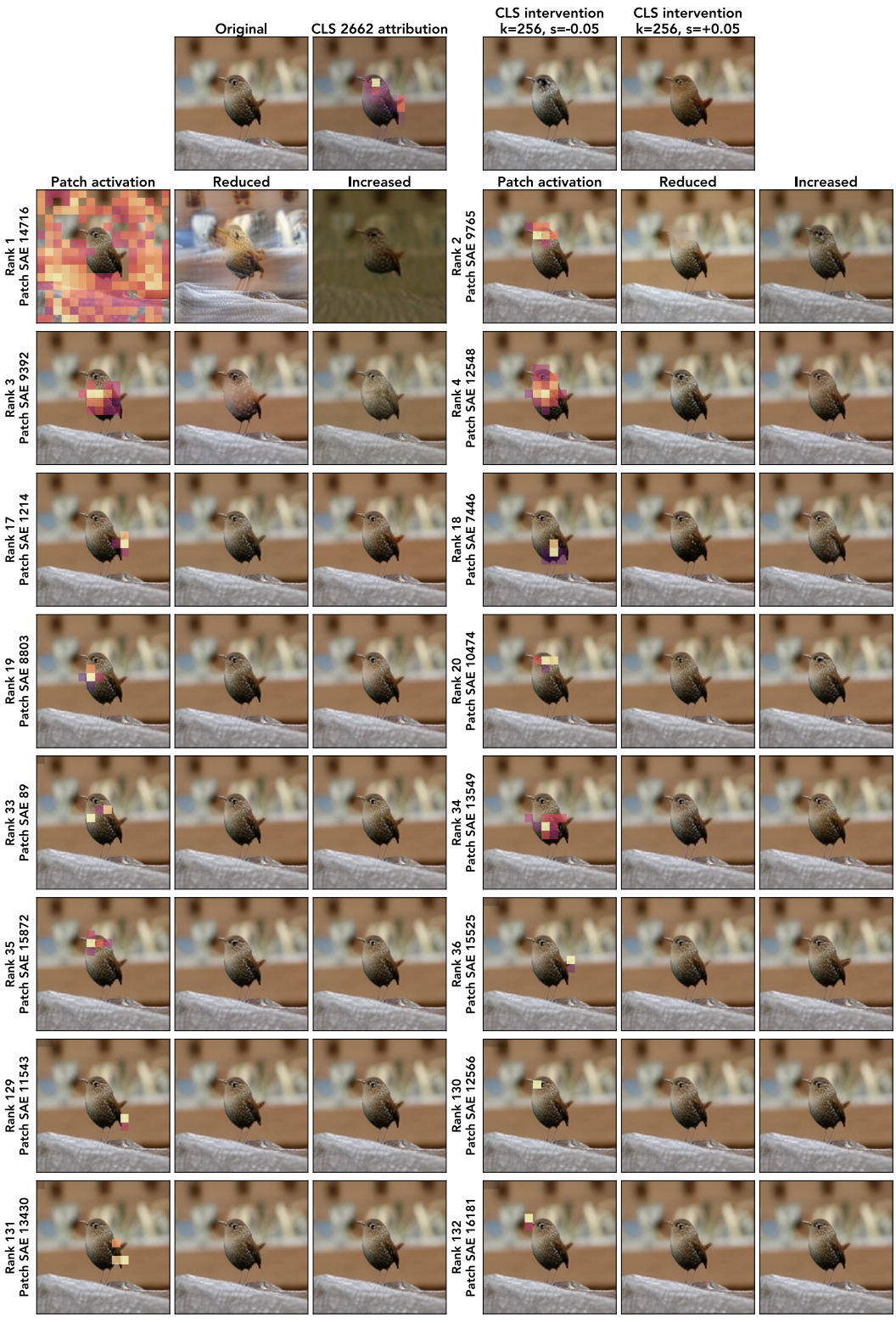


Figure 15: **Attributed Patch-SAE decomposition of `[CLS]` SAE feature** 2662. Highly attributed patch SAE dimensions localize to different image regions and induce distinct visual changes under intervention, revealing the local sparse components that jointly support the global `[CLS]` feature.

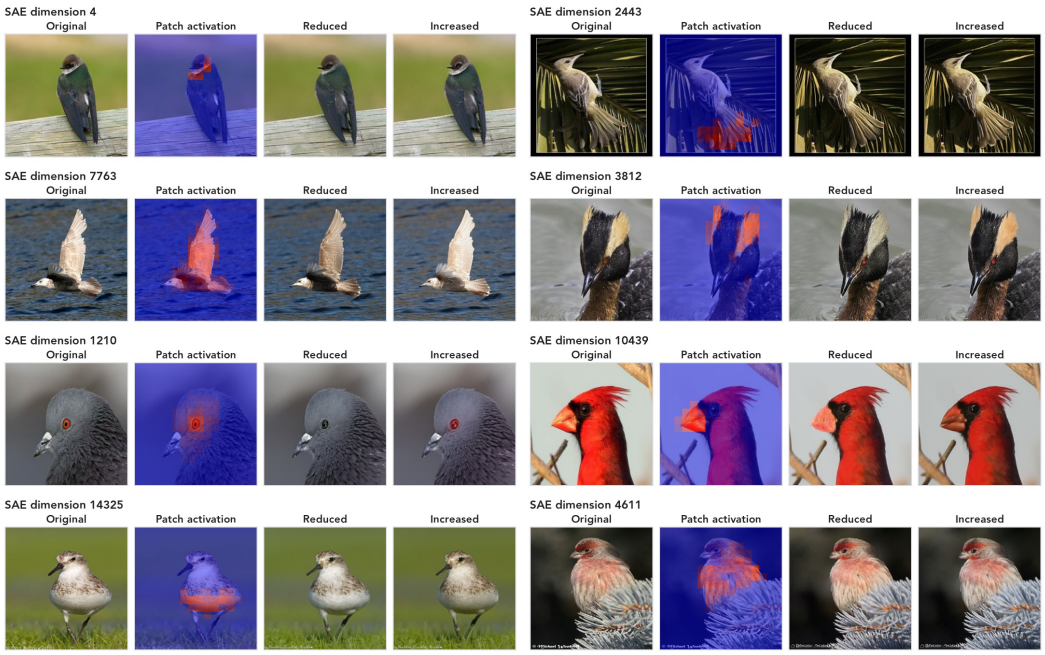


Figure 16: **Intervention reconstruction examples** on patch-SAE features of NABirds extracted with DINOv3.

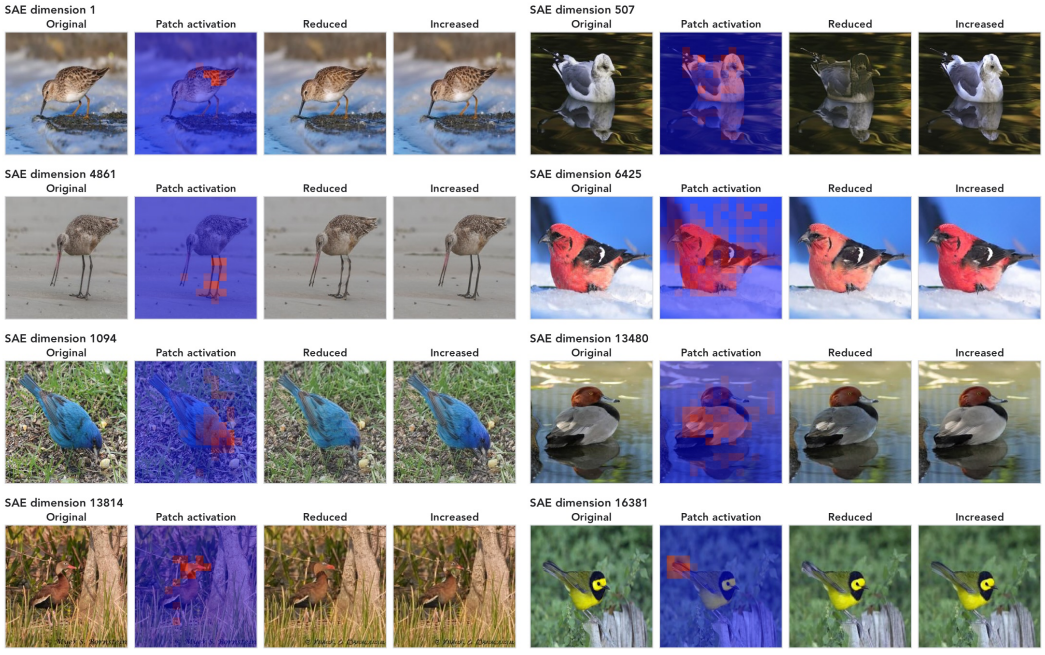


Figure 17: **Intervention reconstruction examples** on patch-SAE features of NABirds extracted with SigLIP 2.

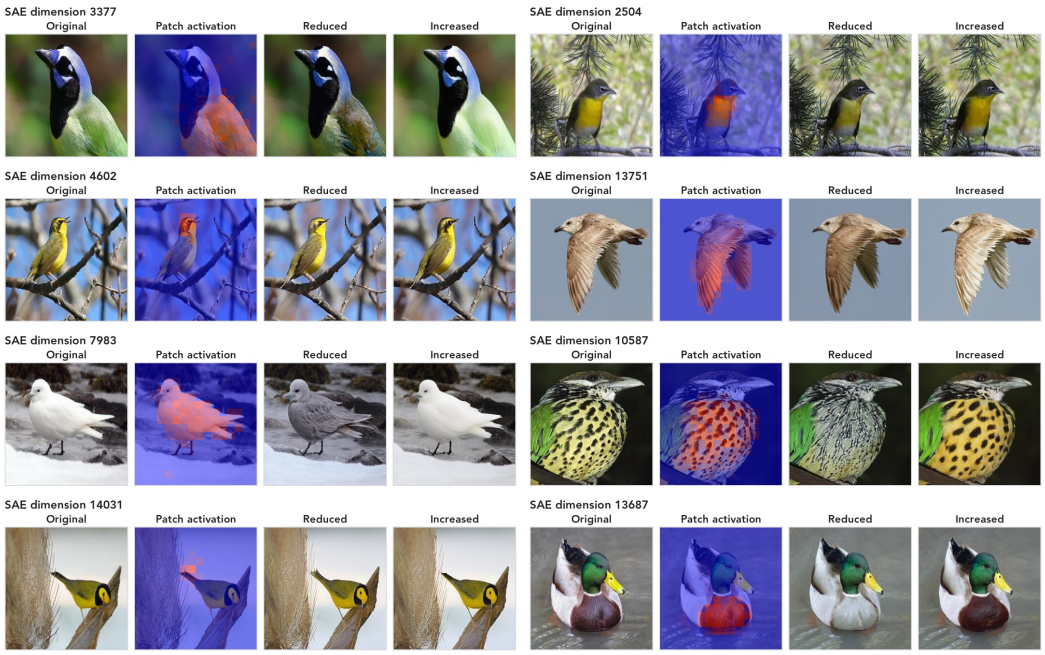


Figure 18: **Intervention reconstruction examples** on patch-SAE features of CUB extracted with DINOv3.

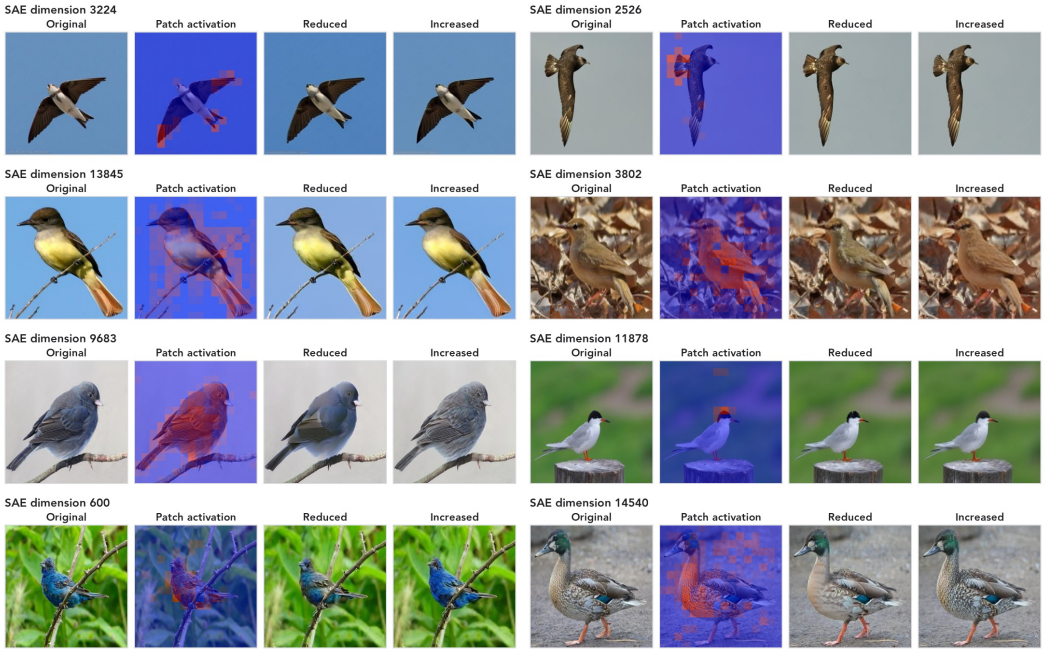


Figure 19: **Intervention reconstruction examples** on patch-SAE features of CUB extracted with SigLIP 2.

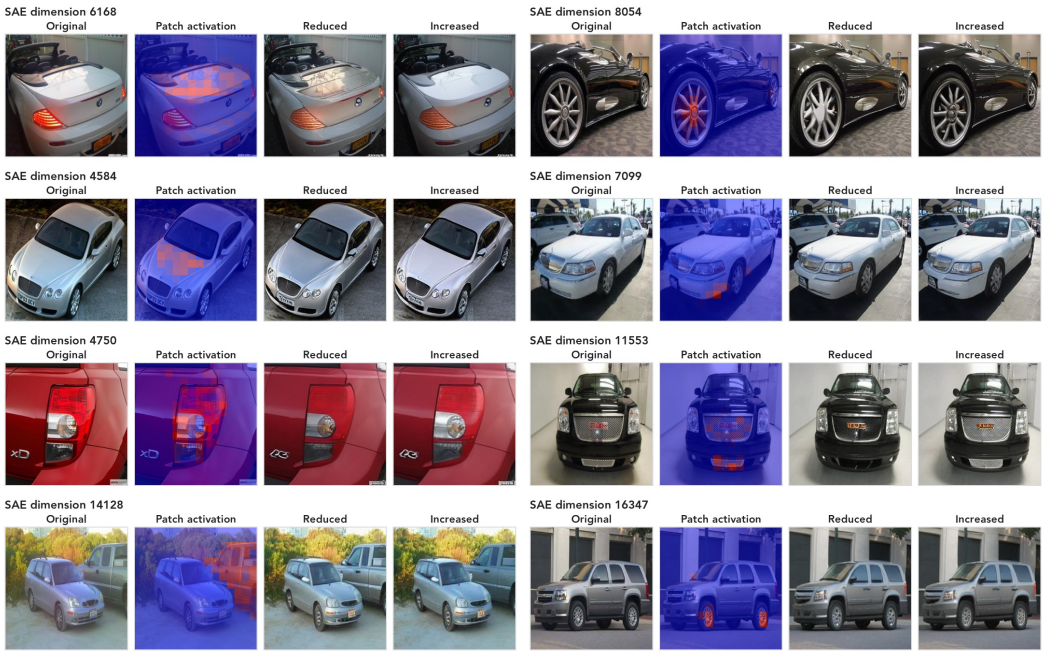


Figure 20: **Intervention reconstruction examples** on patch-SAE features of Cars extracted with DINOv3.



Figure 21: **Intervention reconstruction examples** on patch-SAE features of Cars extracted with SigLIP 2.

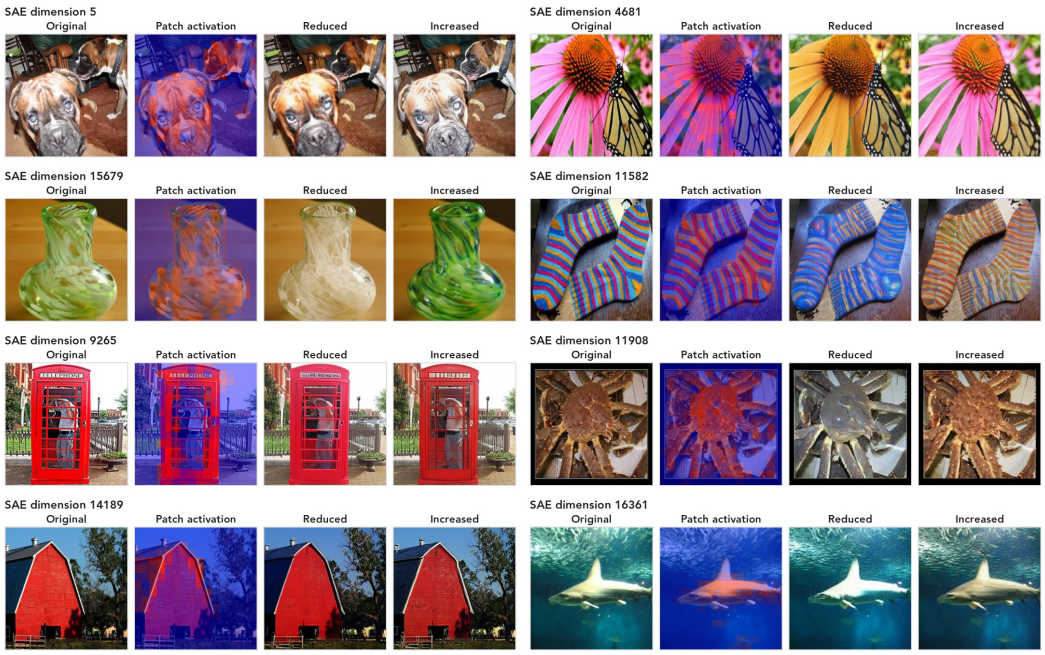


Figure 22: **Intervention reconstruction examples** on patch-SAE features of ImageNet extracted with DINOv3.

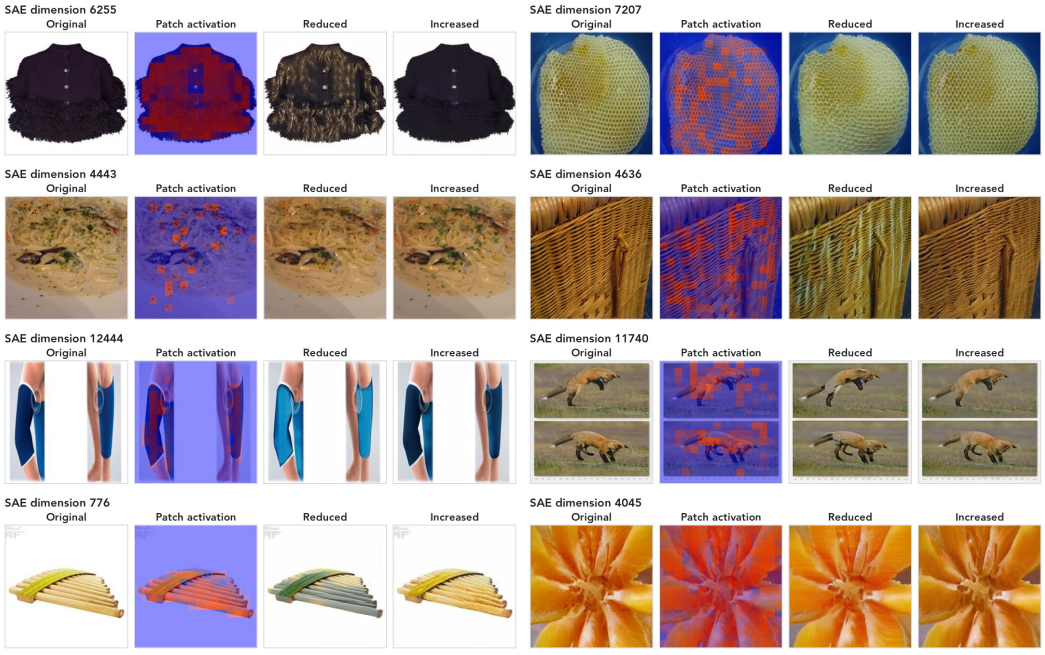


Figure 23: **Intervention reconstruction examples** on patch-SAE features of ImageNet extracted with SigLIP 2.

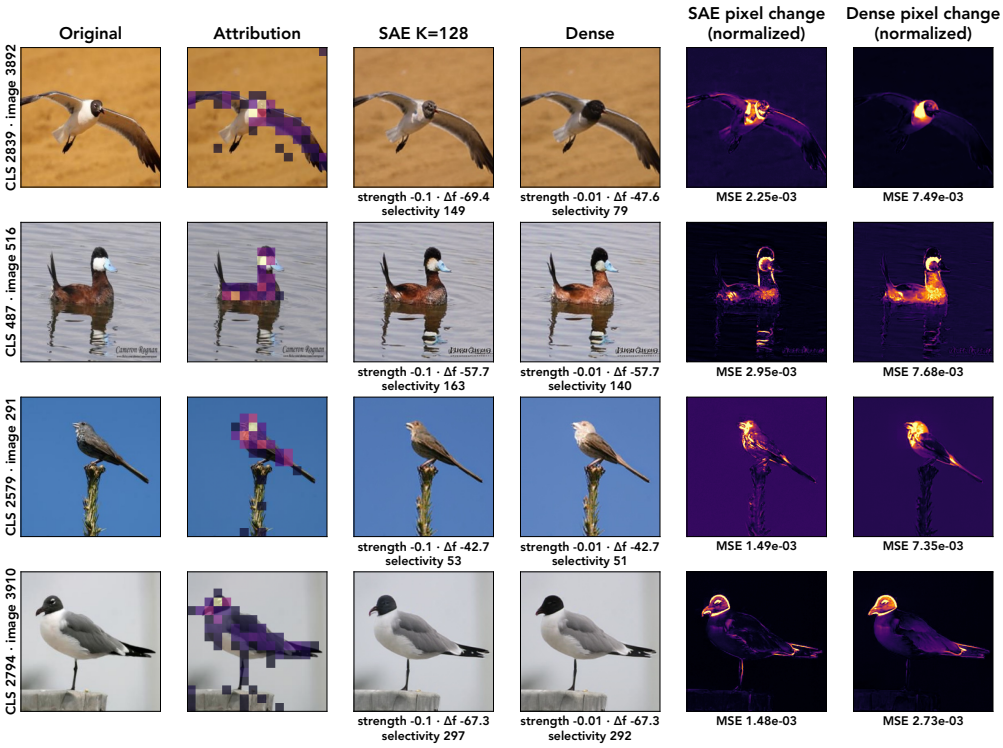


Figure 24: **Examples of suppressed intervention reconstruction** on `[CLS]` SAE features of NABirds extracted with DINOv3.

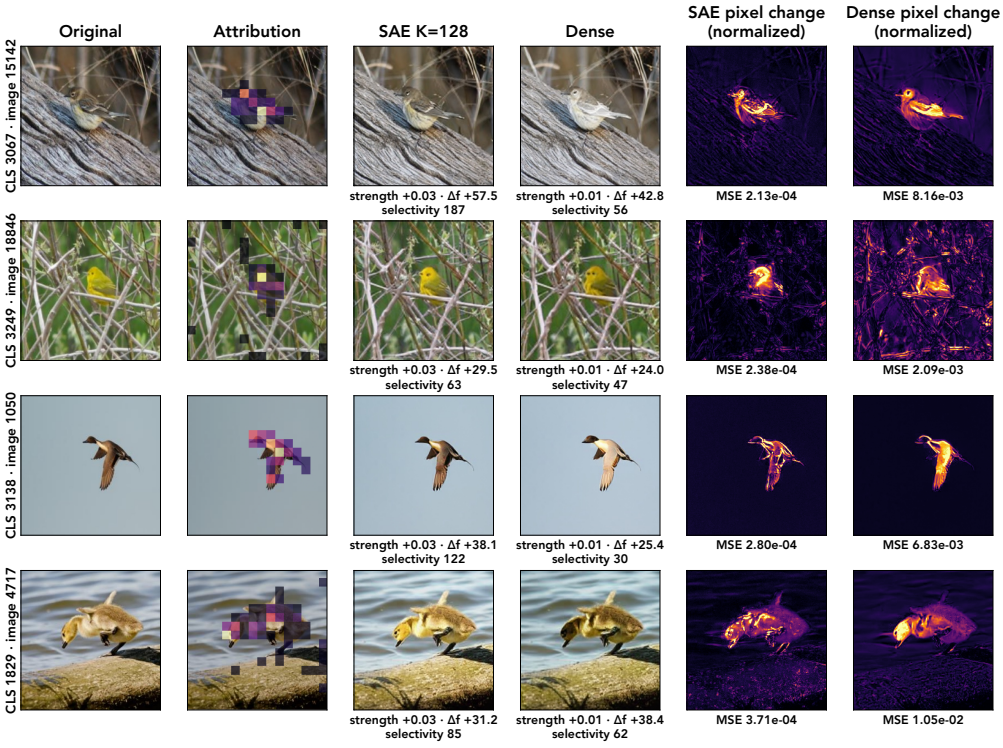


Figure 25: **Examples of amplified intervention reconstruction** on `[CLS]` SAE features of NABirds extracted with DINOv3.

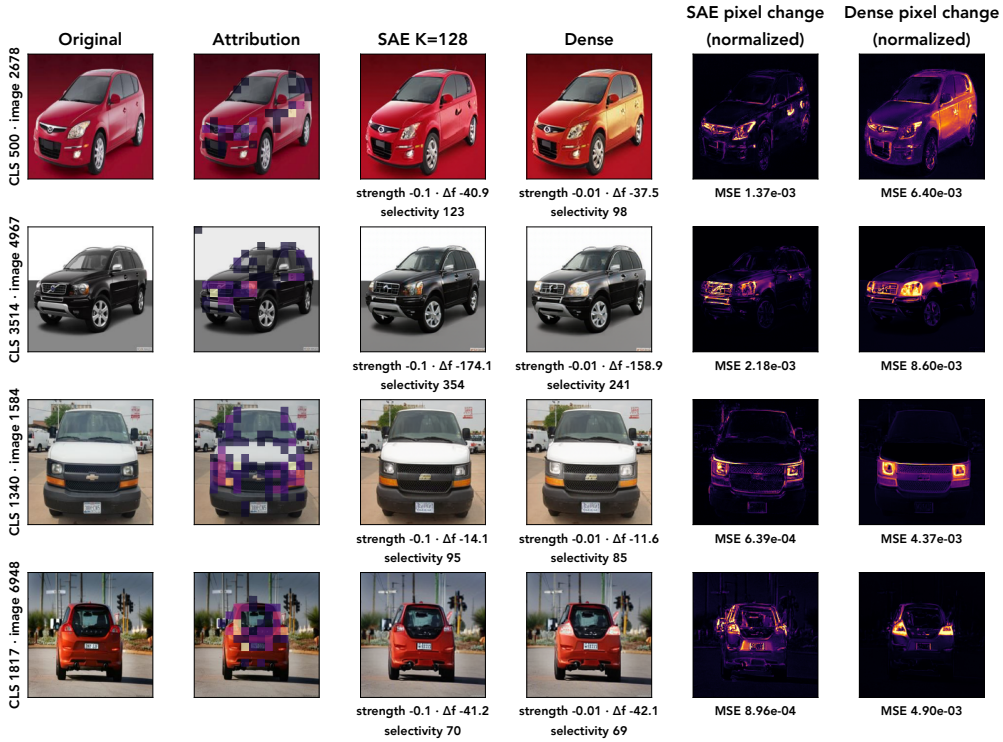


Figure 26: **Examples of suppressed intervention reconstruction** on `[CLS]` SAE features of Cars extracted with DINOv3.

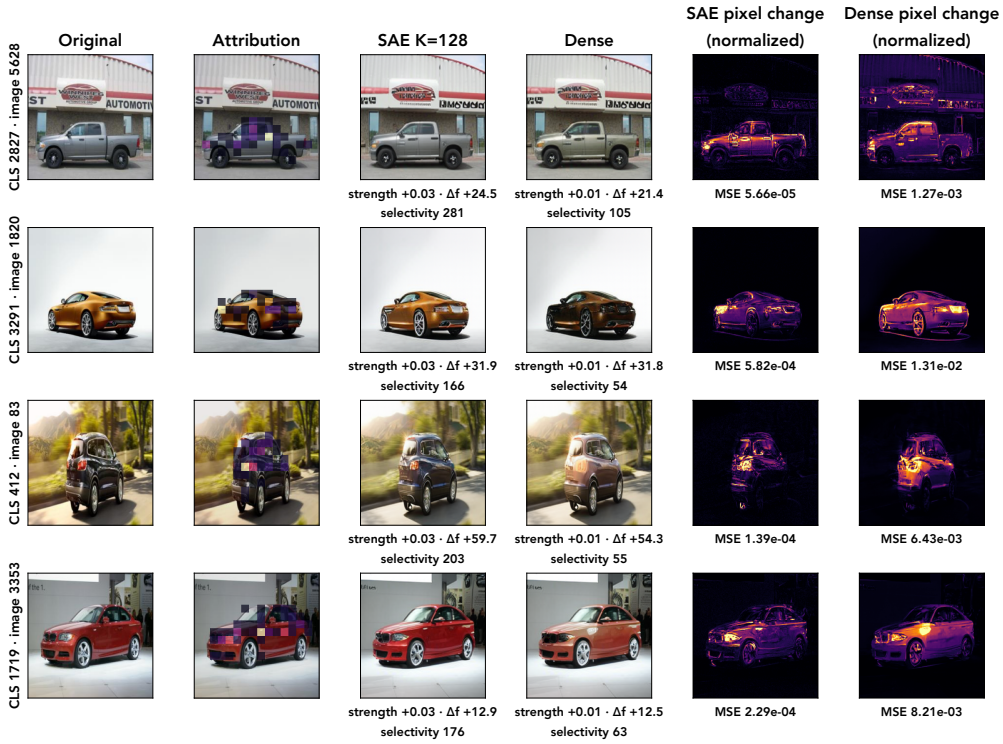
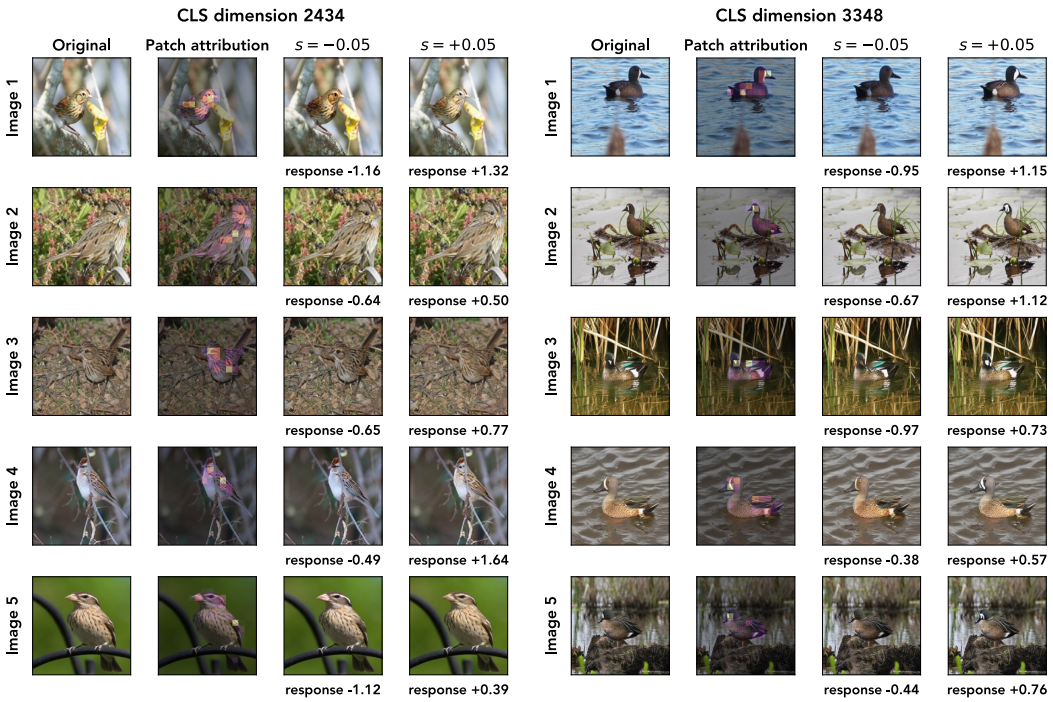


Figure 27: **Examples of amplified intervention reconstruction** on `[CLS]` SAE features of Cars extracted with DINOv3.



(a) (b)

Figure 28: **Intervention reconstruction consistency results** on `[CLS]` features (a) 2434 and (b) 3348 of NABirds extracted with DINOv3.