SPARSE DECOMPOSITION OF CROSS-ATTENTION OV CIRCUITS VIA ISOTROPIC PROBING

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding the semantic influence of text on image generation is essential for interpreting text-to-image diffusion models. Cross-attention provides a direct interface between text and the denoising network, but attention maps reveal where token information is routed rather than what each head writes into the visual feature stream. We focus on the output-value (OV) pathway, where the value and output projections compose into a fixed linear map from token embeddings to feature updates. Motivated by isotropic random measurements in matrix sensing, we introduce IsoProbe, which probes each head's OV operator with isotropic Gaussian inputs and fits a sparse transcoder to the resulting outputs. This data-free procedure yields an overcomplete dictionary of writing directions without requiring text or image samples. The resulting dictionaries approximate OV writes through input-dependent sparse combinations of learned directions. Prompt-based analyses identify concept-associated code subsets within shared heads. Injections, patches, and ablations link selected entries to changes in object, style, and attribute-associated content, providing a means to investigate their functional contributions.

1 INTRODUCTION

Diffusion-based text-to-image models generate diverse and high-fidelity imagery from natural language descriptions (Saharia et al., 2022; Rombach et al., 2022). Understanding this generation process requires tracing how objects, styles, and attributes described in text contribute to visual content. Cross-attention couples text representations with visual features, providing a concrete pathway for examining these semantic contributions.

Each cross-attention head decomposes into two largely independent circuits (Elhage et al., 2021; Bae et al., 2026). The query-key (QK) circuit produces an attention pattern that selects the visual positions to which each text token is routed. The output-value (OV) projection transforms the selected token's representation into a feature vector written into the visual stream at those positions. Existing cross-attention analyses focus largely on the QK side, characterizing how token information is routed (Liu et al., 2024; Tang et al., 2023; Park et al., 2025) and exploiting attention modulation for generation control (Hertz et al., 2023; Chefer et al., 2023; Kim et al., 2023; Rassin et al., 2023). Modifying attention weights, however, alters only the routing of token information. By contrast, the OV projection is the immediate write pathway through which the head contributes text-conditioned content. Head-level concept associations (Park et al., 2025) do not by themselves explain how different concepts recruit components within a shared head. We ask whether learned sparse coordinates provide a useful representation for analyzing and testing concept-associated contributions within a shared head.

The OV pathway has another useful property for mechanistic analysis. For each head, the OV operator is defined as the matrix product of the value and output projections, yielding a fixed linear map from text-encoder embeddings to the visual feature stream. Because this linear transformation is fully determined by the trained parameters, its responses to isotropic random probes can be computed directly. These probes are analogous to the isotropic measurement ensembles used in matrix sensing (Recht et al., 2010; Candès & Plan, 2011). Building on this insight, we approximate each head's OV operator with a sparse transcoder (Dunefsky et al., 2024) whose decoder furnishes an overcomplete dictionary of writing directions. We call this procedure *isotropic probing*. It fits the fixed OV operator over uniformly distributed input directions without text or image samples. After training, we use

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.

prompts to identify concept-associated dictionary elements and interventions to test their contributions to generated content.

Prompt-based analyses reveal concept-associated sparse-code patterns within shared heads, and interventions on selected subsets produce distinct changes in generated content. The same dictionaries support analyses of implicit attribute and safety-relevant signals, including occupation-driven gender bias and adversarial unsafe content. By modifying selected code contributions without changing model parameters or QK projections, we investigate the functional roles of learned writing directions. These results support OV dictionaries as tools for analyzing concept-associated contributions within cross-attention heads.

Our contributions are threefold.

- We learn sparse writing dictionaries directly from fixed cross-attention OV operators using isotropic Gaussian probes, without text or image samples. These dictionaries approximate OV writes through input-dependent sparse combinations of learned directions.
- We analyze concept-associated sparse-code patterns within shared heads through token associations, code overlap, and responses to controlled prompt changes.
- We test the functional contributions of selected dictionary entries through injections, patches, and ablations, linking code-level interventions to changes in object, style, and attribute-associated content.

2 Related Work

Cross-attention analysis in diffusion models. Cross-attention in diffusion models has been studied through attention attribution (Tang et al., 2023), attention-based image editing (Hertz et al., 2023), cross/self-attention behavior analyses in Stable Diffusion-style pipelines (Liu et al., 2024), and attention modulation for generation control (Chefer et al., 2023; Rassin et al., 2023; Kim et al., 2023). Individual cross-attention heads also align with human visual concepts (Park et al., 2025). Most closely related, Bae et al. (2026) analyze cross-attention OV circuits via SVD and localize concepts in low-dimensional spectral subspaces. We extend this OV-circuit perspective by learning sparse writing dictionaries from fixed operators using isotropic probes, and analyzing concept-associated contributions within shared heads through their input-dependent activation patterns.

Circuit analysis. Circuit analysis identifies reusable sub-circuits within transformers and characterizes specific head-level mechanisms (Elhage et al., 2021; Olsson et al., 2022). Recent work extends this approach to vision representations (Gandelsman et al., 2024), and automated circuit discovery enables scalable identification of attention head behaviors (Conmy et al., 2023). We bring this circuit perspective to text-to-image diffusion cross-attention by treating value writing as an explicit linear operator that can be probed and decomposed independently of the attention map.

Sparse dictionary learning. The linear representation hypothesis (Park et al., 2024), together with work on superposition (Elhage et al., 2022) and polysemanticity (Scherlis et al., 2022), motivates dictionary-based decompositions of neural network internals, and sparse autoencoders realize this idea with overcomplete representations whose features can be inspected at scale (Huben et al., 2024; Bricken et al., 2023; Templeton et al., 2024; Gao et al., 2025). Transcoders extend the approach to approximate transformations rather than reconstruct activations (Dunefsky et al., 2024), and sparse feature circuits combine dictionary learning with causal analysis (Marks et al., 2025). In diffusion models, SAEs trained on generated activations have been used to identify semantic features, erase concepts, and mitigate bias (Surkov et al., 2025; Kim et al., 2025; Cywiński & Deja, 2025; Tian et al., 2025; Shi et al., 2025). We instead train a sparse *transcoder* that approximates an explicit OV operator using probe-response pairs from the operator, yielding a sparse dictionary for reading and testing head-wise OV writes rather than a reconstruction of generation-time activations.

3 Method

For each cross-attention head, we construct a sparse transcoder that approximates the fixed OV operator, yielding an overcomplete dictionary of writing directions whose entries support both

inspection and intervention. Section 3.1 formalizes the OV operator, Section 3.2 develops the transcoder dictionary, and Section 3.3 presents the training objective. Section 3.4 specifies the sparse-code analysis and intervention protocols. Figure 1 provides an overview.

3.1 Cross-Attention OV Writing Operators

Consider a cross-attention layer with H heads, visual features $\mathbf{z}$, and text embeddings $\mathbf{c}$. All vectors are treated as row vectors. The output of head h can be written as

$$\mathrm{Head}^{(h)}(\mathbf{z},\mathbf{c}) = \underbrace{\mathbf{A}^{(h)}(\mathbf{z},\mathbf{c})}_{\text{QK, } \textit{where}}\ \underbrace{\mathbf{c}\,\mathbf{W}_{\mathrm{V}}^{(h)}\mathbf{W}_{\mathrm{O}}^{(h)}}_{\text{OV, } \textit{what}}, \tag{1}$$

where $\mathbf{A}^{(h)}(\mathbf{z},\mathbf{c}) = \mathrm{softmax}\big(\mathbf{z}\mathbf{W}_{\mathrm{Q}}^{(h)}(\mathbf{c}\mathbf{W}_{\mathrm{K}}^{(h)})^{\top}/\sqrt{d_h}\big)$ is the attention matrix. The QK circuit determines where token information is routed, while the OV circuit performs the linear transformation that determines what is written into the feature stream (Elhage et al., 2021), illustrated in Figure 1a. Since the OV transformation is a fixed composition of the value and output projections and independent of the spatial attention pattern, it can be analyzed statically. We define the head's OV operator as

$$\mathbf{W}_{\mathrm{OV}}^{(h)} := \mathbf{W}_{\mathrm{V}}^{(h)}\mathbf{W}_{\mathrm{O}}^{(h)} \in \mathbb{R}^{d_c\times d}, \tag{2}$$

where $\mathbf{W}_{\mathrm{V}}^{(h)} \in \mathbb{R}^{d_c\times d_h}$ and $\mathbf{W}_{\mathrm{O}}^{(h)} \in \mathbb{R}^{d_h\times d}$ are the value and output projections for head h, d_c is the text encoder embedding dimension, d is the visual feature dimension, and d_h is the head dimension. For a token-level hidden state $x \in \mathbb{R}^{d_c}$ from the text encoder, the written vector is $x\mathbf{W}_{\mathrm{OV}}^{(h)} \in \mathbb{R}^{d}$. Given any chosen probe x, its OV output $x\mathbf{W}_{\mathrm{OV}}^{(h)}$ can be computed directly.

3.2 Sparse Transcoders as OV Writing Dictionaries

We seek a decomposition that can describe each OV write as a sparse combination of learned writing directions in the visual feature space. For this reason, we approximate each OV operator with an overcomplete sparse transcoder, following sparse dictionary approaches for identifying interpretable directions in neural representations (Bricken et al., 2023; Gao et al., 2025). Spectral decompositions express the operator as a sum of rank-one linear maps (Bae et al., 2026), whereas our sparse transcoder uses input-dependent subsets of learned decoder directions. Suppressing a code affects only writes where it is active, motivating a selectivity comparison with spectral ablation (Section 4.2).

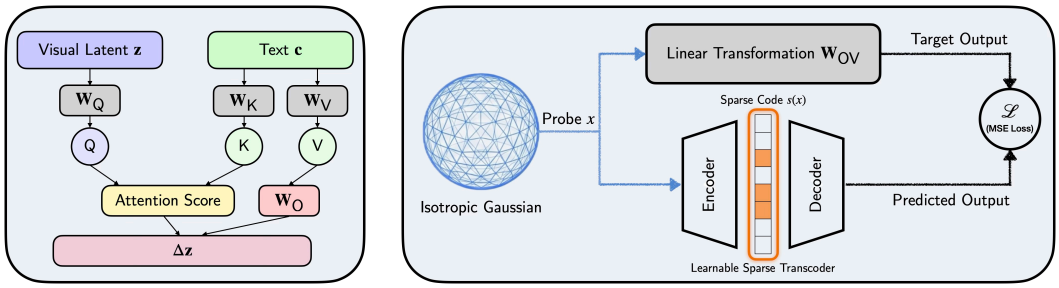
For each head, we train a sparse transcoder T_θ that approximates the OV writing operator $x \mapsto x\mathbf{W}_{\mathrm{OV}}^{(h)}$ through an overcomplete sparse code, illustrated in Figure 1b. The encoder produces a sparse code, and the decoder maps it to an approximate OV write. Each dictionary element is a learned decoder direction, while its corresponding sparse-code entry specifies the direction's activation for a given input. Unlike a sparse autoencoder, which reconstructs its own input (Huben et al., 2024; Bricken et al., 2023), a transcoder approximates a transformation from the input to a distinct target activation (Dunefsky et al., 2024). In our setting, the target is the head's OV write, so the decoder dictionary resides in the operator's output space. The decoder directions are learned to approximate the operator's exact responses under the sparse coding constraint.

The transcoder consists of an encoder, a sparsifying nonlinearity, and a linear decoder

$$s(x) = \mathrm{TopK}\big(\mathrm{ReLU}(x\mathbf{W}_{\mathrm{enc}}),\, k\big), \qquad T_\theta(x) = s(x)\,\mathbf{W}_{\mathrm{dec}}, \tag{3}$$

where $\mathbf{W}_{\mathrm{enc}} \in \mathbb{R}^{d_c\times M}$, $\mathbf{W}_{\mathrm{dec}} \in \mathbb{R}^{M\times d}$, and M is the dictionary size. The encoder maps an embedding $x \in \mathbb{R}^{d_c}$ to M pre-activations. ReLU first enforces nonnegativity, and TopK then retains the k largest nonnegative activations and zeroes the rest, yielding a code $s(x)$ with at most k nonzero entries and $k \ll M$. The decoder then writes the predicted output as a nonnegative combination of at most k active rows of $\mathbf{W}_{\mathrm{dec}}$, each a vector in the d-dimensional OV output space. The encoder thus selects *which* writing directions a token activates, while the decoder defines *what* those directions write into the feature stream.

We use an overcomplete dictionary (e.g., $M = 8\,d_c$), so the M decoder rows can represent more distinct writing directions than the operator's input dimension, and different active subsets of $s(x)$ can express qualitatively different OV writes. These dictionary elements are learned components introduced by the transcoder, not literal units in the original diffusion model.



(a) Cross-attention circuit.

(b) Training on OV probes.

Figure 1: (a) A cross-attention head uses QK projections to form attention scores and the output-value path to write an update Δz to the visual stream. (b) For each head, isotropic Gaussian probes are evaluated under $\mathbf{W}_{\mathrm{OV}}$. A sparse transcoder receives the same probes and learns sparse codes whose decoded outputs match the OV outputs under an MSE loss.

3.3 Training with Isotropic Probes

Since $\mathbf{W}_{\mathrm{OV}}^{(h)}$ is an explicit linear operator, each chosen probe x gives an exact probe-response pair $(x, x\mathbf{W}_{\mathrm{OV}}^{(h)})$. Isotropic Gaussian probes sample input directions without favoring those prevalent in a particular text corpus. We use the resulting input-output pairs to learn a sparse approximation of the fixed OV operator. We sample probe embeddings x from an isotropic Gaussian $\mathcal{N}(0, I_{d_c})$ and compute exact targets $y = x\mathbf{W}_{\mathrm{OV}}^{(h)}$. We then minimize mean squared error

$$\mathcal{L}(\theta) = \mathbb{E}_{x\sim\mathcal{N}(0,I_{d_c})}\left[\|T_\theta(x) - x\mathbf{W}_{\mathrm{OV}}^{(h)}\|^2\right]. \tag{4}$$

This training uses no text data. Text prompts enter only after the dictionary has been learned, when we inspect or edit the sparse codes they activate. Transcoders trained on caption embeddings provide a useful comparison, but they fit the same objective only on directions emphasized by a particular caption distribution. Appendix B gives empirical diagnostics for this difference, and Appendix L formalizes the measurement geometry, including a conditional cone-level coverage result under directional concentration of OV writes (Appendix L.5).

3.4 Sparse-Code Analysis and Intervention Tests

After training, the sparse representation supports two complementary uses, *descriptive analysis* and *intervention tests*. For descriptive analysis, the dictionary elements selected by the support of $s(x)$ indicate which learned writing directions are used by a token or prompt component. Concept-associated directions can be identified either by comparing sparse codes for a base prompt and a concept-augmented prompt, or by measuring alignment between decoder rows and reference directions in the OV output space.

In our main image interventions, we replace the head's OV write with the transcoder output and edit its sparse code,

$$\tilde{v} = \mathrm{edit}\big(s(x)\big)\,\mathbf{W}_{\mathrm{dec}}, \tag{5}$$

where $\mathrm{edit}(\cdot)$ either *suppresses* selected entries of the top-k sparse code by setting them to zero, or *injects* additional activation into selected entries of $s(x)$. Injection is additive, so the edited code can contain both the original token entries and the injected entries before decoding. These interventions test selected code contributions in the transcoder approximation without modifying model parameters or QK computation at the intervention site. Appendix B.3 evaluates no-edit substitution fidelity. Appendix E.3.2 separately tests object interventions by subtracting selected decoded contributions from the original attention output.

4 Experimental Results

We first evaluate the fidelity and write selectivity of the sparse OV approximations (Section 4.2). We then examine concept-associated contributions within shared heads (Section 4.3) and investigate

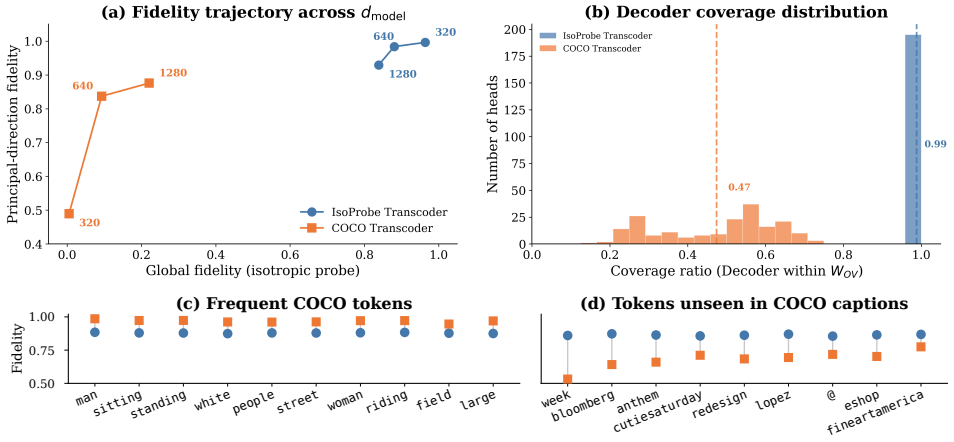


Figure 2: OV-operator fidelity and decoder coverage for Stable Diffusion v2.1 cross-attention heads. In (a), 320, 640, and 1280 denote the U-Net visual feature dimensions d_{model}, corresponding to the OV output dimension d. Gaussian-trained transcoders achieve higher cosine fidelity to $x\mathbf{W}_{\text{OV}}$ under global and principal-direction probes, while also concentrating decoder energy within the OV output subspace. (c, d) Token-level fidelity (cosine, mean over 195 heads) on real CLIP token states. (c) Ten frequent COCO content tokens, in 1,000 COCO captions outside the training set. (d) Selected COCO-unseen tokens from P4D and Ring-A-Bell K=38.

their functional roles through sparse-code interventions (Section 4.4). Finally, we examine implicit attribute and safety signals (Section 4.5).

4.1 Experimental Setup

We instantiate the framework on Stable Diffusion v2.1 and study all 195 cross-attention heads across the U-Net. The CLIP text encoder (Radford et al., 2021) produces $d_c = 1024$-dimensional token embeddings. For each head, we extract the OV matrix $\mathbf{W}_{\text{OV}}^{(h)}$ and train a sparse transcoder with dictionary size $M = 8192$ ($8\times$ overcomplete) and fixed sparsity $k = 16$. Sensitivity to the sparse budget k and dictionary size M is reported in Appendix J. To compare probe distributions for the same transcoder objective, we train two variants with identical optimization budgets. The *IsoProbe* transcoder draws probes from an isotropic Gaussian, $x \sim \mathcal{N}(0, I_{d_c})$ (Section 3.3), while the *COCO* transcoder uses CLIP embeddings of 29K COCO (Lin et al., 2014) captions. Implementation details are provided in Appendix A. Unless otherwise stated, mechanistic analyses use the IsoProbe transcoder.

4.2 OV Operator Fidelity and Write Selectivity

Reading OV writes through sparse codes requires high operator fidelity. We evaluate how well the transcoder reproduces the head's OV writes on held-out probes.

We measure fidelity by the cosine similarity between the transcoder output $T_\theta(x)$ and the ground-truth write $x\mathbf{W}_{\text{OV}}$, using two probe families. *Global fidelity* averages this similarity over isotropically sampled input directions. *Principal-direction fidelity* instead averages it over probes aligned with the top left singular vectors of $\mathbf{W}_{\text{OV}}$, which correspond to input directions with the largest write magnitudes for unit-norm inputs. *Decoder coverage* is a complementary diagnostic, $\text{cov}(\mathbf{W}_{\text{dec}}) = \|\mathbf{W}_{\text{dec}} P_S\|_F^2 / \|\mathbf{W}_{\text{dec}}\|_F^2$ with P_S the projection onto $\text{Row}(\mathbf{W}_{\text{OV}})$. Values near one mean the dictionary writes inside the operator's range rather than spending capacity on off-subspace directions.

On Stable Diffusion v2.1, IsoProbe transcoders achieve high OV-write fidelity, with principal-direction fidelity above 0.9 in Figure 2a and decoder coverage $\text{cov} \approx 0.99$ in Figure 2b. COCO-trained transcoders place less than half their decoder energy in the OV output subspace ($\text{cov} \approx 0.47$) and have lower global and principal-direction fidelity.

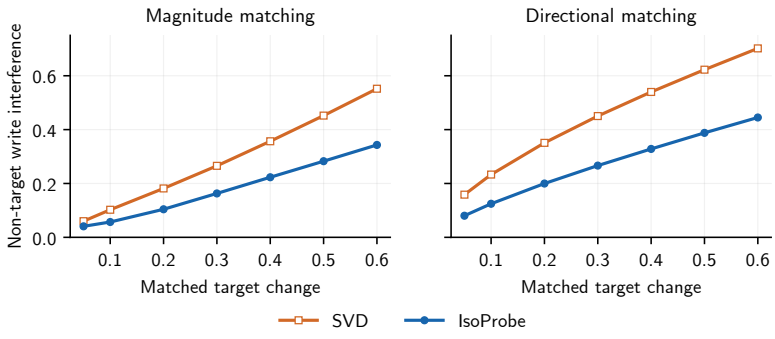


Figure 3: Mean non-target OV-write interference under target changes matched by relative magnitude (left) or reduction along the original write direction (right). Lower is better.



(a) Baseline. (b) Zero apple. (c) Zero pear. (d) Per-head selected sparse-code count.

Figure 4: Concept-associated sparse-code interventions within the top-10 shared OV heads where apple- and pear-associated sparse-code entries contribute most strongly. (a–c) Zeroing the selected apple or pear sparse-code entries in these heads collapses the targeted object slot toward the competing fruit morphology, while largely preserving the non-target object and the scene layout. (d) Per-head selected sparse-code count, separately for apple-specific and pear-specific selected entries.

This gap is consistent with differences in probe coverage. Concentrated probe distributions can leave the transcoder weakly constrained along input directions with little training support (Appendix L). Isotropic probing provides broad directional coverage and yields higher operator fidelity and decoder coverage in these experiments. To examine how this gap affects image generation, Appendix B.3 compares the images produced by the two transcoders on identical prompts.

We also evaluate OV-write fidelity on contextual CLIP token representations (Figure 2c,d). The COCO transcoder achieves higher fidelity on the frequent COCO tokens shown in panel (c), while IsoProbe achieves higher fidelity on the selected COCO-unseen tokens in panel (d) and maintains similar fidelity across the two sets of examples.

We next compare sparse-code suppression with SVD-based spectral ablation (Bae et al., 2026) to evaluate write selectivity. SVD ablates rank-one linear components of the same operator, whereas IsoProbe removes contributions only from selected codes active for the input. Because their contributions differ in magnitude, we match the target write change rather than the number of removed components. The two criteria match relative change magnitude or reduction along the original write direction. We then fix each method's selected components and strengths and measure changes in other prompts' writes. IsoProbe yields lower mean non-target interference across the evaluated range under both criteria (Figure 3). Appendix G.1 details the evaluation settings, while Appendix G.2 examines image-level responses separately.

4.3 Concept-Associated Contributions within Shared Heads

Prior work reports that cross-attention heads and their attention-map position patterns align with human-specified visual concepts (Park et al., 2025). A shared head can contribute to multiple concepts. We examine how these contributions are represented by concept-associated sparse-code patterns and test their functional roles through interventions on selected entries.



(a) Style injection. (b) Cross-concept patching.

Figure 5: Sparse-code style injection and object patching. (a) Adding style entries to the cat-token code on the top-20 style-relevant heads transfers style while preserving the scene. (b) Replacing dog entries with cat entries gives a partial hybrid at $N_{\text{total}} = 100$ and a category shift at $N_{\text{total}} = 500$. N_{total} counts selected entries across all heads.

Code overlap describes which dictionary entries different concepts recruit, whereas interventions test what those entries contribute to generated content. A small overlap indicates distinct usage of the learned representation, but functional differences must be assessed through the model's response to interventions. We therefore compare both the selected code sets and the effects of suppressing them within the same shared heads.

The apple/pear prompt provides a controlled test case ("a painting of a single red apple on the left and a single green pear on the right"). At the head level, apple and pear recruit almost the same set of OV heads, and the top-10 shared heads each carry entries associated with both concepts. At the feature level, however, the raw sparse-code activations on the two concept tokens exhibit low overlap. For each prompt component (apple, pear, red, green, painting), we identify the top sparse features per head at the component's token, and define apple-exclusive features as those in apple's top set but not in red, green, pear, or painting (full protocol in Appendix E.1). As shown in Figure 4d, this yields 125 apple- and 120 pear-exclusive entries within these shared heads. We zero the selected entries across all text-token positions in the conditional branch of these ten heads. As shown in Figure 4a–c, ablating the apple-exclusive entries collapses the apple slot toward pear morphology, and ablating the pear-exclusive entries collapses the pear slot toward apple morphology, with the non-target object and the scene layout largely preserved. Both interventions act on the same heads and all text-token positions, with the target determined by the selected code subset. Their distinct effects therefore provide a functional test of whether different code subsets within shared heads contribute to different objects.

Appendix E.2 tests apple/pear code reuse across prompts and sampling settings. Appendix E.3 extends to other object pairs with fixed-code evaluation in held-out contexts (E.3.1) and interventions retaining the original attention output (E.3.2). Keeping selected codes fixed tests whether their functional roles generalize beyond the selection prompt. Retaining the original attention output tests whether those effects also occur without replacing the model's writes by transcoder reconstructions. Further analyses show low code overlap across 39 concepts (Appendix E.4) and greater code sensitivity to concept replacement than to spatial wording (Appendix C).

4.4 Functional Tests through Sparse-Code Interventions

We extend this analysis to adding, replacing, and progressively suppressing codes through style injection, object patching, and style ablation. Injection adds activation into selected entries of $s(x)$, with style entries drawn from base/style prompt differences. Patching replaces source-concept entries with target-concept entries on the same token slot. Ablation sets selected sparse-code entries to zero before decoding.

As shown in Figure 5a, style injection into the cat-token code on the top-20 style-relevant heads turns the baseline cat into Van Gogh, cartoon, and sketch variants while preserving layout. Cross-concept patching pooled across all 195 heads produces progressively stronger dog-to-cat transfer in Figure 5b, from a partial hybrid at $N_{\text{total}} = 100$ to a complete category shift at $N_{\text{total}} = 500$. Progressive style ablation in Figure 6 removes target signatures (Van Gogh brushwork, pixel-art blocks, cyberpunk lighting, retro styling) while retaining scene layout.

These interventions support a functional interpretation of the selected dictionary entries as contributors to object- and style-related OV writes within cross-attention heads. The same ablation pattern is

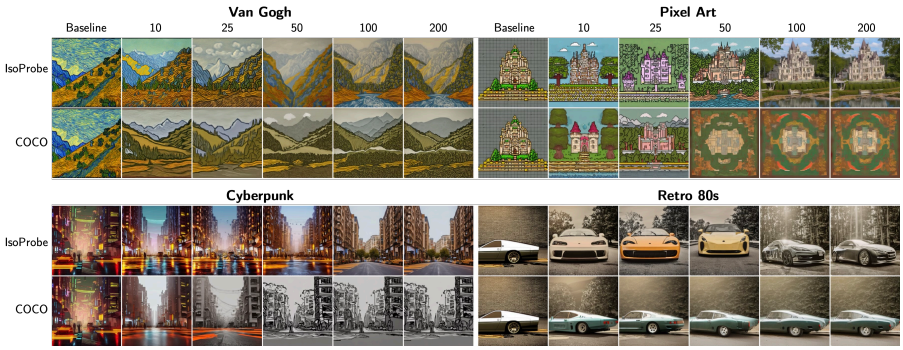


Figure 6: Progressive style suppression with IsoProbe (upper row) and COCO (lower row) for each style. Columns show the baseline followed by suppression of $N \in \{10, 25, 50, 100, 200\}$ selected dictionary entries per head.

Table 1: Occupation-level OV gender z-scores relative to the neutral-control baseline. Negative scores (top rows) and positive scores (bottom rows) indicate projections below and above the neutral-control mean, corresponding to the male and female sides of the reference direction, respectively.

Occupation	z	Occupation	z	Occupation	z
police officer	-5.25	mechanic	-4.73	actor	-4.62
musician	-4.05	politician	-3.72	manager	-3.48
firefighter	-3.03	CEO	-2.27	doctor	-1.46
dietitian	$+1.97$	model	$+2.03$	nurse	$+2.15$
secretary	$+2.45$	fashion model	$+2.50$	stylist	$+2.78$

observed in Stable Diffusion v1.4, SDXL (Podell et al., 2024), and PixArt-α (Chen et al., 2024) (Appendix I).

4.5 Causal Interventions on Implicit Attribute and Safety Signals

Beyond explicit object and style concepts, we apply the OV dictionary to two implicit signals that cross-attention heads write and that bear directly on auditing diffusion behavior: occupation-driven gender bias and adversarial unsafe content. Intervening on the corresponding sparse-code entries causally shifts the generated content without modifying the query-key projections or other model parameters.

Tracing gender bias. Prior work reports that gender bias in text-to-image diffusion can originate in the text encoder and propagate through the diffusion process (Wu et al., 2024; Wang & Specia, 2024; Han et al., 2025). The OV write dictionary exposes this propagated attribute signal. A gender reference is extracted as a female-minus-male sparse-code pattern from paired female and male prompts and decoded into a gender direction in the OV write. Each occupation's OV write is scored by its alignment with this direction, normalized by a gender-neutral baseline of control words to give a z-score (Appendix F). Table 1 reports deviations toward the male ($z < 0$) and female ($z > 0$) sides relative to the neutral-control baseline.

As shown in Figure 7 for $\alpha = \pm 0.5$ versus $\alpha = 0$, editing along the gender-reference direction with coefficient α shifts apparent gender presentation, with $\alpha < 0$ pushing toward the male side and $\alpha > 0$ toward the female side. Appendix F traces this signal through three aligned views per occupation in Figure 23, showing that dictionary-element overlap with the gender reference, OV-space z-score, and CLIP-classified apparent gender of the generated image all move together. Occupation-level gender bias therefore passes through sparse OV write patterns that can be edited at inference time.

Safety-relevant directions. We extend the intervention analysis to safety-related OV writes. We rank dictionary entries by alignment with a safety-reference sparse-code pattern constructed from



Figure 7: Intervention along the gender-reference pattern. Columns show $\alpha = -0.5$, baseline ($\alpha = 0$), and $\alpha = +0.5$ for each occupation. Negative α pushes apparent gender presentation toward the male side of the pattern and positive α pushes it toward the female side.

Table 2: Safety-direction suppression as an intervention check. All methods in this table are evaluated on Stable Diffusion v2.1. ASR (%, ↓) on I2P, MMA, P4D, and UnLearnDiffAtk, with FID relative to unmodified SDv2.1 outputs and CLIP for text–image alignment on COCO-1K. Baselines group into training-based, closed-form editing, and inference-time guidance families. Full table in Appendix H. TC denotes transcoder.

Method	I2P ↓	MMA ↓	P4D ↓	UnLearnDiffAtk ↓	FID ↓	CLIP ↑
SDv2.1	14.95	22.20	52.68	32.39	–	31.40
ESD (Gandikota et al., 2023)	8.17	2.10	32.68	17.61	38.08	30.85
RECE (Gong et al., 2024)	3.08	0.00	6.83	6.34	43.17	30.69
SLD-Strong (Schramowski et al., 2023)	2.12	1.90	12.68	3.52	41.93	30.68
SAFREE (Yoon et al., 2025)	5.41	2.90	27.32	13.38	41.51	30.90
IsoProbe TC ($N = 0$)	14.63	20.40	50.24	26.06	40.02	31.07
IsoProbe TC ($N = 25$)	4.45	5.90	28.29	8.45	41.12	30.93
IsoProbe TC ($N = 50$)	1.80	1.90	12.20	3.52	43.52	30.35

unsafe-reference prompts and matched benign prompts, using a selection set disjoint from the adversarial evaluation prompts. Suppressing the selected entries reduces unsafe-content generation relative to the no-suppression reconstruction, supporting their functional association with the targeted content. Table 2 summarizes this intervention check. The $N = 0$ row reports transcoder substitution without code suppression. Increasing suppression from $N = 0$ to $N = 50$ reduces ASR from 14.63% to 1.80% on I2P and from 20.40% to 1.90% on MMA. Detailed benchmark comparisons and image-level results are reported in Appendix H.

5 Conclusion

Attention maps show where cross-attention routes token information, but the OV operator determines what each head writes. We introduced IsoProbe, which trains sparse transcoders on isotropic Gaussian probes of fixed OV operators without text or image samples. The resulting dictionaries approximate OV writes through input-dependent sparse combinations of learned directions and reveal concept-associated code patterns within shared heads. At matched target write changes, sparse-code interventions produce lower non-target OV-write interference than spectral ablations in our evaluation. Interventions on selected entries link these patterns to changes in object, style, and attribute-associated content, providing a means to investigate their functional contributions. Reconstruction error limits the fidelity of our main substitution-based image interventions. Additional object interventions retain the original attention output and produce concept-associated changes, although non-target appearance is not fully preserved. Future work includes reducing reconstruction error, integrating OV analysis with QK routing, and extending selective sparse-code interventions to joint-attention architectures, where text and image streams interact. IsoProbe thus provides an operator-based approach to analyzing and testing the contributions of text to image generation.

AI USE STATEMENT

We used large language models (LLMs) to review the grammar, wording, and logical flow of manuscript drafts written by the authors. For the theoretical analysis in Appendix L, LLMs were used to review the assumptions and logical steps of propositions and proofs formulated by the authors. The authors verified the final propositions and proofs. Agentic coding tools were used to extend code originally written by the authors, and the authors reviewed the resulting code. All experimental outputs were produced by the models and methods evaluated in this paper, and all reported numerical results were obtained from executed experimental code. The authors take responsibility for all content of this paper, including text and code developed with LLM assistance.

ETHICS STATEMENT

This work analyzes and edits the content that cross-attention heads write during text-to-image generation, and two aspects warrant comment. Section 4.5 measures how each occupation's OV write aligns with a gender-reference pattern and intervenes along that direction. The reported z-scores describe alignment within the studied model and are not normative claims about the occupations. The sign convention carries no evaluative meaning. The gender-reference prompts and occupation templates used to construct the pattern are listed in Appendix F so that the construction can be inspected. The same substrate that audits attribute and safety signals at inference time can also insert them, so deployment should pair OV-level controls with model-level access and monitoring. Our safety evaluation uses established public adversarial benchmarks and reports aggregate attack success rates (Appendix H).

REPRODUCIBILITY STATEMENT

The method and code-level intervention procedures are described in Section 3. Appendix A provides implementation details, including transcoder training and image-generation settings. Experiment-specific protocols and evaluation metrics are reported in the corresponding appendices. Appendix L provides the assumptions and proofs for the theoretical analysis. We provide the experimental code in the supplementary materials.

REFERENCES

Jun-Hyun Bae, Wonyong Jo, Jaehyup Lee, and Heechul Jung. Mechanistic dissection of cross-attention subspaces in text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 19657–19665, 2026. doi: 10.1609/AAAI.V40I24.39046.

Praneeth Bedapudi. Nudenet: Neural nets for nudity classification, detection and selective censoring, 2019. URL `https://github.com/notAI-tech/NudeNet`.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Emmanuel J. Candès and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inf. Theory*, 57(4):2342–2359, 2011. doi: 10.1109/TIT.2011.2111771.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Trans. Graph.*, 42(4), 2023. doi: 10.1145/3592116.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James T. Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations*, 2024.

Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. In *Proceedings of the International Conference on Machine Learning*, pp. 8468–8486, 2024.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, 2023.

Bartosz Cywiński and Kamil Deja. SAeUron: Interpretable concept unlearning in diffusion models with sparse autoencoders. In *Proceedings of the International Conference on Machine Learning*, volume 267, pp. 11738–11775, 2025.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable LLM feature circuits. In *Advances in Neural Information Processing Systems*, 2024.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. In *International Conference on Learning Representations*, 2024.

Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2426–2436, 2023.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5111–5120, 2024.

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

Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. In *European Conference on Computer Vision*, pp. 73–88, 2024. doi: 10.1007/978-3-031-73668-1_5.

Boyu Han, Qianqian Xu, Shilong Bao, Zhiyong Yang, Kangli Zi, and Qingming Huang. Lightfair: Towards an efficient alternative for fair t2i diffusion via debiasing pre-trained text encoders. In *Advances in Neural Information Processing Systems*, 2025.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.595.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

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

Dahye Kim, Xavier Thomas, and Deepti Ghadiyaram. Revelio: Interpreting and leveraging semantic information in diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4659–4669, 2025.

Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7667–7677, 2023. doi: 10.1109/ICCV51070.2023.00708.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22691–22702, 2023.

Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12006–12016, 2024. doi: 10.1109/CVPR52733.2024.01141.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.

Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7817–7826, 2024. doi: 10.1109/CVPR52733.2024.00747.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. MACE: mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6430–6440, 2024. doi: 10.1109/CVPR52733.2024.00615.

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

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

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection, 2024. URL https://arxiv.org/abs/2306.09683.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

Jungwon Park, Jungmin Ko, Dongnam Byun, Jangwon Suh, and Wonjong Rhee. Cross-attention head position patterns can align with human visual concepts in text-to-image generative models. In *International Conference on Learning Representations*, 2025.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the International Conference on Machine Learning*, pp. 39643–39666, 2024.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. In *Advances in Neural Information Processing Systems*, 2023.

Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, 2010. doi: 10.1137/070697835.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022.

Adam Scherlis, Kshitij Sachan, Adam S Jermyn, Joe Benton, and Buck Shlegeris. Polysemanticity and capacity in neural networks. *arXiv preprint arXiv:2210.01892*, 2022.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22522–22531, 2023. doi: 10.1109/CVPR52729.2023.02157.

Yingdong Shi, Changming Li, Yifan Wang, Yongxiang Zhao, Anqi Pang, Sibei Yang, Jingyi Yu, and Kan Ren. Dissecting and mitigating diffusion bias via mechanistic interpretability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8192–8202, 2025. doi: 10.1109/CVPR52734.2025.00767.

Viacheslav Surkov, Chris Wendler, Antonio Mari, Mikhail Terekhov, Justin Deschenaux, Robert West, Caglar Gulcehre, and David Bau. One-step is enough: Sparse autoencoders for text-to-image diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: interpreting stable diffusion using cross attention. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 5644–5659, 2023. doi: 10.18653/V1/2023.ACL-LONG.310.

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

Zhihua Tian, Sirun Nan, Ming Xu, Shengfang Zhai, Wenjie Qu, Jian Liu, Ruoxi Jia, and Jiaheng Zhang. Sparse autoencoder as a zero-shot classifier for concept erasing in text-to-image diffusion models. *arXiv preprint arXiv:2503.09446*, 2025.

Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-A-Bell! how reliable are concept removal methods for diffusion models? In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/b5b4d92374323c53c24bbbc8ee0e715c-Abstract-Conference.html.

Guorun Wang and Lucia Specia. Moesd: Mixture of experts stable diffusion to mitigate gender bias. *arXiv preprint arXiv:2407.11002*, 2024.

Yankun Wu, Yuta Nakashima, and Noa Garcia. Stable diffusion exposed: Gender bias from prompt to image. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 1648–1659, 2024. doi: 10.1609/AIES.V7I1.31754.

Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7737–7746, 2024. doi: 10.1109/CVPR52733.2024.00739.

Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. SAFREE: training-free and adaptive guard for safe text-to-image and video generation. In *International Conference on Learning Representations*, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images ... for now. In *European Conference on Computer Vision*, pp. 385–403, 2024. doi: 10.1007/978-3-031-72998-0_22.

APPENDIX CONTENTS

A Implementation Details **16**
A.1 Generation Settings . . . 16
A.2 Edit Application Protocol . . . 16

B Operator Fidelity Diagnostics **16**
B.1 COCO Embedding Geometry . . . 16
B.2 Controlled Synthetic Operator Study . . . 16
B.3 No-Edit Substitution Fidelity . . . 18
B.4 Global OV Approximation with the Full COCO Caption Pool . . . 18
B.5 Operator Approximation across Architectures . . . 19

C OV-Code Sensitivity to Position and Concept **19**

D Token Selectivity Across Heads **20**
D.1 Token Associations of Dictionary Elements . . . 20
D.2 Van Gogh-Associated Top Codes . . . 21

E Concept-Associated Contributions within Shared Heads **23**
E.1 Sparse-Code Selection Protocol . . . 23
E.2 Prompt and Sampling Variants . . . 23
E.3 Generalization Across Concept Pairs . . . 24
E.4 Sparse-Code Overlap on a Concept Panel . . . 26
E.5 Image-level Validation of Sparse-Code Interventions . . . 27

F Gender-Association Methodology **32**

G Comparison with Spectral Subspace Concept Removal **34**
G.1 Matched OV-Write Interventions . . . 34
G.2 Qualitative Comparison under Increasing Ablation . . . 35

H Safety Intervention Comparison **35**
H.1 Additional Sparse-Dictionary Comparison on SDv1.4 . . . 39

I Cross-Architecture Concept Removal **39**

J Sparse Budget and Dictionary Size in PixArt-α **40**

K Exploratory OV Analysis in SD3 Joint Attention **41**

L Theoretical Analysis **43**
L.1 Setup and Notation . . . 43
L.2 Gaussian Probing Gives a Uniform-Average Directional Objective . . . 44
L.3 Anisotropic Probing Weakly Constrains Low-Variance Directions . . . 44
L.4 Decoder Coverage and Off-Subspace Edit Artifacts . . . 45
L.5 Cone-Level Coverage Under Directional Concentration . . . 46

A Implementation Details

For Stable Diffusion v2.1, we train one sparse transcoder for each of the 195 U-Net cross-attention heads. Each transcoder consists of a linear encoder, a ReLU followed by TopK that retains the largest k nonnegative activations, and a linear decoder. Each transcoder targets the per-head OV product $\mathbf{W}_{\mathrm{OV}}^{(h)} = \mathbf{W}_{\mathrm{V}}^{(h)}\mathbf{W}_{\mathrm{O}}^{(h)}$ alone.

We compare IsoProbe and COCO probe sources under the same optimization budget. The COCO variant uses 30K COCO captions, with 1K held out as a validation set. Training uses the $29\mathrm{K} \times 77$ token embeddings obtained by passing the remaining 29K captions through the Stable Diffusion v2.1 tokenizer and text encoder. The IsoProbe variant uses the same number of probes sampled independently from $\mathcal{N}(0, I_{d_c})$. Both variants are trained with AdamW using learning rate 10^{-3}, weight decay 10^{-5}, batch size 512, and 48,800 optimization steps per head. Training all 195 transcoders takes approximately 2 hours and 24 minutes on $8\times$ NVIDIA A100 40 GB GPUs. Since each head's OV matrix is independent of the others, transcoder training is fully parallelizable across heads.

A.1 Generation Settings

We use the public `stable-diffusion-2-1-base` checkpoint with its default PNDM scheduler. We sample 25 inference steps with classifier-free guidance scale 7.5 at 512×512 resolution.

A.2 Edit Application Protocol

N counts the dictionary entries selected for suppression per head, whereas k limits the entries active at each token before editing. The sparse code is obtained from the unmodified text-encoder embedding, and the same code change is applied at all denoising timesteps. The edit is therefore fully determined by the prompt and does not adapt to the noisy latent. All edits act only on the OV write, so we do not modify the QK weights, and the direct intervention site is the OV content pathway rather than the QK routing pathway.

B Operator Fidelity Diagnostics

We examine the directional concentration of COCO embeddings (Section B.1), test its effect on a controlled synthetic operator (Section B.2), and evaluate image generation under no-edit transcoder substitution (Section B.3).

Isotropic Gaussian probes $x \sim \mathcal{N}(0, I_{d_c})$ have root-mean-square norm $\sqrt{d_c} \approx 32.0$. Empirical CLIP text-embedding norms across COCO and the I2P, P4D, and MMA safety benchmarks fall in the 33.3–33.5 range, within 5% of this value. Since our fidelity metric is cosine-based (Section 4.2), this small scale offset does not affect operator-direction coverage.

B.1 COCO Embedding Geometry

Our observational baseline trains the transcoder on the COCO-29K text-embedding distribution. As a directional diagnostic, we project COCO-29K text embeddings, isotropic Gaussian probes, and out-of-distribution prompts onto the top principal components of the COCO embedding set. The visualization is qualitative and illustrates the directional concentration of an observational text dataset.

Figure 8 shows that COCO embeddings cluster in narrow regions of direction space, while isotropic Gaussian probes cover the input space more uniformly. The next subsection tests how a transcoder approximating a linear operator behaves when its training probes exhibit this anisotropy.

B.2 Controlled Synthetic Operator Study

We isolate the effect of probe anisotropy in a controlled setting where the target operator is fully specified. We fix a target matrix $\mathbf{W} \in \mathbb{R}^{64\times64}$ and train transcoders to approximate $x \mapsto x\mathbf{W}$. We compare four probe distributions. The first is an isotropic Gaussian $\mathcal{N}(0, I_{64})$. The remaining three are k-dominant anisotropic Gaussians ($k \in \{8, 16, 32\}$) whose diagonal covariances concentrate

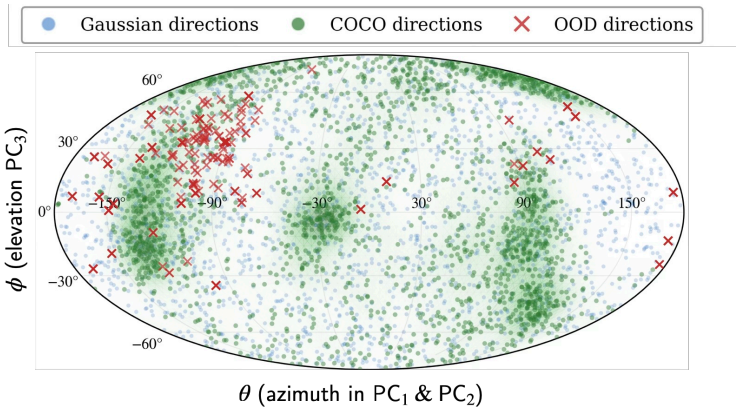


Figure 8: Directional distribution of text embeddings projected onto the top-3 principal components of COCO captions. COCO embeddings (green) cluster in narrow regions of direction space, while isotropic Gaussian probes (blue) provide broader coverage. Out-of-distribution prompt directions (red) are shown for comparison.

variance in the first k dimensions, mimicking the directional concentration typical of finite text-embedding sets. Each covariance is normalized to $\mathrm{tr}(\Sigma) = 64$ to match the total variance of the isotropic case. We evaluate every trained transcoder on all four distributions using a 4×4 cross-evaluation matrix.

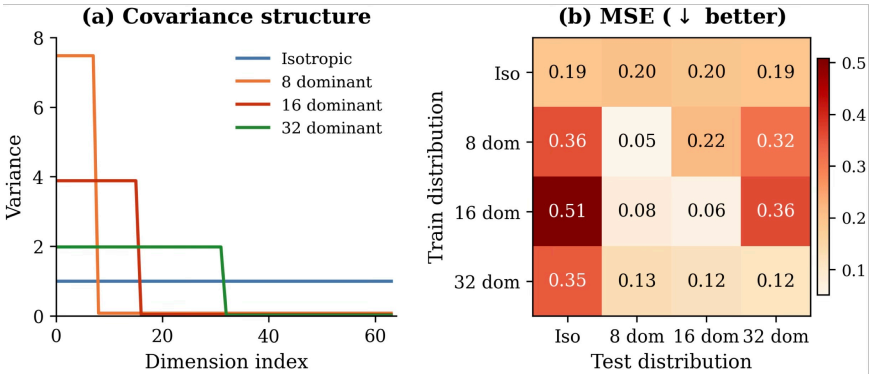


Figure 9: Controlled study on a synthetic operator. The left panel shows the diagonal covariance structure of the four probe distributions, where the k-dominant distributions concentrate variance in a small number of dimensions. The right panel reports cross-distribution test MSE for transcoders approximating $x \mapsto x\mathbf{W}$ with $\mathbf{W} \in \mathbb{R}^{64\times64}$, with rows indicating training distribution and columns indicating test distribution. Only the isotropic transcoder generalizes uniformly across all test conditions.

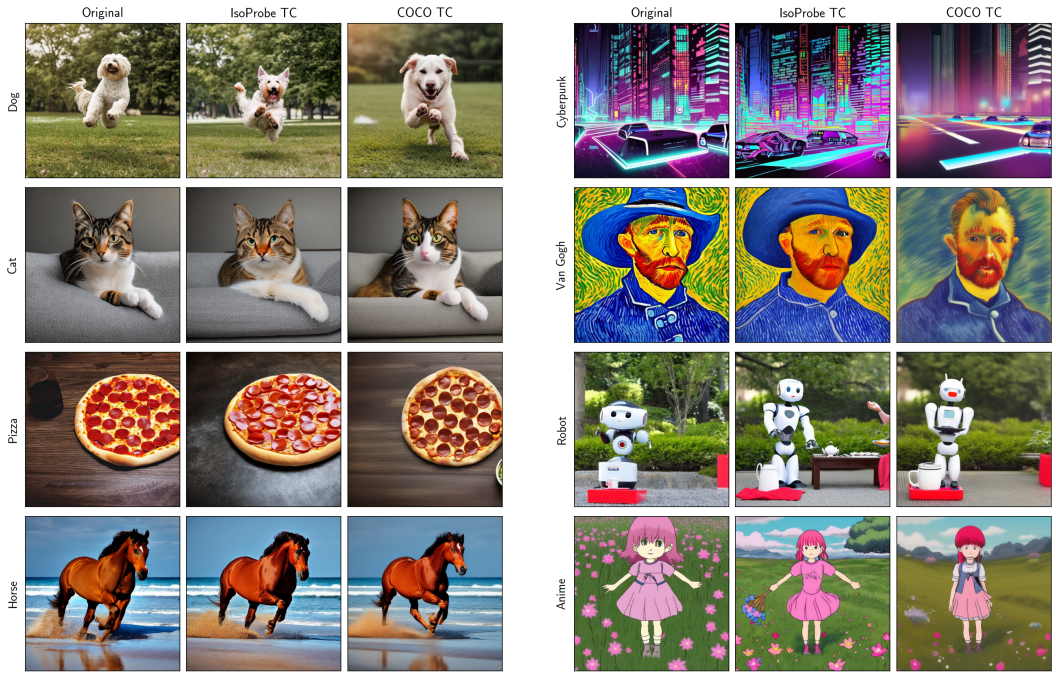
Figure 9 shows that transcoders trained on anisotropic distributions achieve low error within their training support but degrade on held-out directions, while the isotropic transcoder maintains uniformly low error across all test conditions. Faithful global operator approximation therefore benefits from directional coverage over the full input space. On actual OV operators in Figure 2 of the main paper, IsoProbe transcoders achieve higher fidelity under the evaluated global and principal-direction probes, while COCO transcoders allocate more decoder energy outside the OV output subspace. Formal statements appear in Appendix L, including the anisotropic underconstraint result and the decoder-coverage bound on off-subspace artifacts.

B.3 No-Edit Substitution Fidelity

We evaluate how well the original model's generation is preserved under no-edit substitution by replacing $x\mathbf{W}_{\mathrm{OV}}$ with $T_\theta(x)$ for all 195 cross-attention heads and comparing the resulting images with the unmodified model outputs.

Figure 10a shows in-distribution prompts with common COCO-style objects (dog, cat, pizza, horse), and Figure 10b shows out-of-distribution prompts with artistic styles and unusual concepts (cyberpunk scene, Van Gogh portrait, robot, anime character). The COCO 29K-caption training set frequently names the IID concepts (dog 1,339, cat 1,525, pizza 430, horse 865), whereas the OOD prompts test concepts that are rarely or never named in these captions (cyberpunk 0, Van Gogh 0, robot 3, anime 2). In the displayed in-distribution examples, both transcoders retain the main objects and overall scene content. In the displayed out-of-distribution examples, particularly the Van Gogh and cyberpunk scenes, IsoProbe retains the original outputs' distinctive stylistic features more closely than COCO, consistent with the operator fidelity gap reported in Section 4.2.

On COCO-1K, no-edit substitution yields mean CLIP scores close to the unmodified model, with 31.07 for IsoProbe and 31.06 for COCO compared with 31.40 for SDv2.1 (Tables 12 and 13). The corresponding LPIPS distances to the unmodified outputs are 0.253 and 0.249 (Table 13, $N = 0$).



(a) In-distribution prompts.

(b) Out-of-distribution prompts.

Figure 10: No-edit substitution fidelity. Each row is one prompt, with the unmodified model output (Original) on the left and full 195-head substitutions $x\mathbf{W}_{\mathrm{OV}} \to T_\theta(x)$ for the IsoProbe and COCO transcoders on the right. (a) Both transcoders retain the main objects and overall scene content in the displayed in-distribution examples. (b) IsoProbe retains distinctive stylistic features more closely than COCO, particularly in the Van Gogh and cyberpunk examples.

B.4 Global OV Approximation with the Full COCO Caption Pool

We evaluate global OV approximation using the full COCO (Lin et al., 2014) caption pool. A transcoder trained on all 616,767 captions reaches global fidelity 0.553, compared with 0.883 for IsoProbe on the same heads (Figure 11). The gap persists with the full caption pool under the same optimization budget.

Experimental settings. We use SDv2.1 with $M = 8{,}192$ and $k = 16$. Both conditions use the same 24 sampled heads, with eight from each input/output-shape group. Global fidelity is the mean cosine similarity to the original OV write on 20,000 isotropic Gaussian inputs per head. The three group means receive weights 25, 50, and 120 according to their full-model head populations, giving a population-weighted estimate from the sample.

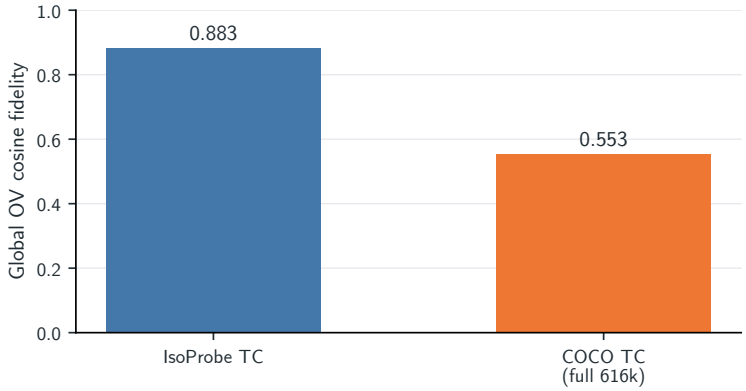


Figure 11: Global OV fidelity for IsoProbe and the full COCO caption baseline on SDv2.1 under the same optimization budget. Both conditions use the same 24 sampled heads, with means weighted by the full head populations of the shape groups.

Together with the concentration of caption-input directions in Appendix B.1, this result supports the role of input-direction coverage in OV approximation. Isotropic probes provide this coverage without a caption corpus.

B.5 Operator Approximation across Architectures

We test the same data-free OV fitting procedure on U-Net-based SDXL (Podell et al., 2024) and DiT-based PixArt-α (Chen et al., 2024). Across all cross-attention heads, IsoProbe reaches fidelity 0.848 and 0.868, respectively, compared with 0.054 and 0.696 for COCO-30K training (Table 3). IsoProbe has higher global fidelity than the caption baseline on every head in both models. These measurements extend the operator-approximation result to both architectures. The corresponding image-level suppression examples are presented separately in Appendix I.

Experimental settings. We train separate dictionaries with $M = 8{,}192$ and $k = 16$ for all 1,300 cross-attention heads in SDXL and all 448 in PixArt-α under both training distributions. Each head is evaluated on 20,000 fresh Gaussian inputs, and fidelity is averaged uniformly over heads.

Table 3: Global OV fidelity across architectures. Fidelity averages all cross-attention heads in each model.

Model	IsoProbe fidelity ↑	COCO-30K fidelity ↑
SDXL	0.848	0.054
PixArt-α	0.868	0.696

C OV-Code Sensitivity to Position and Concept

We compare how OV sparse codes respond to changes in spatial wording and concept replacement.

Protocol. We construct 20 object pairs (A, B). Each pair yields two prompts that swap the stated left-right positions, "a photo of left A and right B, plain white background" and "a photo of left B

and right A, plain white background." Placing the position word before the noun lets the noun hidden state carry the position context, and keeps the noun token indices fixed across the two prompts so that the noun-token sparse codes are directly comparable. At each noun token, we collect sparse codes from all 195 heads using $M = 8192$ and $k = 16$. Active coordinates are indexed by head and dictionary element.

Same concept, new position compares each object's codes across left and right positions. *Same position, new concept* compares the two objects' codes at each stated position. Each condition yields 40 comparisons from the 20 pairs.

Overlap metrics. Let S_1, S_2 denote the active-coordinate sets and s_1, s_2 the corresponding sparse activation vectors. *Shared fraction*, $|S_1 \cap S_2|/|S_1|$, measures support overlap. *Weighted Jaccard*, $\sum_i \min(s_1[i], s_2[i]) / \sum_i \max(s_1[i], s_2[i])$, also accounts for activation magnitudes. *Score cosine*, $\langle s_1, s_2 \rangle / (\|s_1\| \|s_2\|)$, compares the activation vectors. All coordinates index head–element pairs, with zero activation outside the support. We average each metric over the comparisons in each condition.

Results. All three metrics agree (Table 4). When the stated position of the same object changes between left and right, sparse OV codes retain about 86% of their active coordinates and a cosine of 0.92. When the object concept is replaced at the same stated position, retention drops to about 27% and cosine to 0.25.

Table 4: OV-code sensitivity to spatial wording and concept replacement. Sparse OV write codes overlap substantially more under changes in spatial wording (same concept, new position) than under concept replacement (same position, new concept). $n = 40$ comparisons per condition (mean $\pm$ std).

Condition	Shared fraction	Weighted Jaccard	Score cosine
Same concept, new position	0.861 ± 0.014	0.755 ± 0.021	0.916 ± 0.011
Same position, new concept	0.272 ± 0.013	0.125 ± 0.008	0.254 ± 0.016

D Token Selectivity Across Heads

D.1 Token Associations of Dictionary Elements

Section 4.2 showed that the IsoProbe dictionary captures the OV write at the operator level. We now examine which tokens are associated with the activation of individual dictionary elements across a 500-prompt panel. An *active direction* is a dictionary element that enters the top-k on at least one prompt in this panel, and non-active directions are excluded from the selectivity computation. For each active direction, we compute the dominant-token selectivity, the fraction of prompts for which a single token accounts for the majority of that element's activation. The number of active dictionary elements varies per head. Dictionary elements are sorted in decreasing order of dominant-token selectivity.

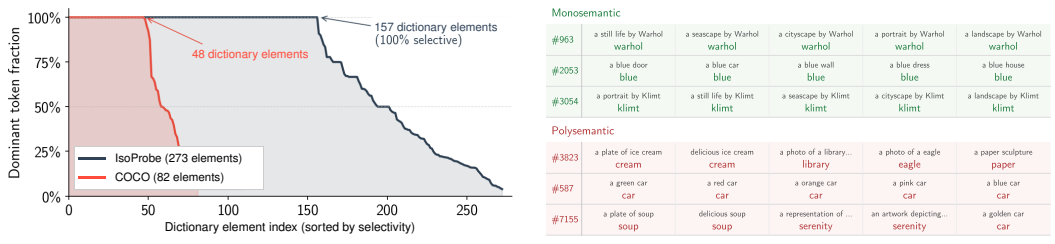


Figure 12: Dictionary-element selectivity for a single cross-attention head (head 2, first down-block). The left panel shows dominant-token selectivity curves for IsoProbe and COCO transcoders, sorted by selectivity. The right panel shows the top-5 activating prompts for representative token-specific and mixed dictionary elements in the IsoProbe transcoder.

On the representative head in Figure 12, 157 of 273 active IsoProbe elements (57.5%) attain 100% dominant-token selectivity, compared with 48 of 82 COCO elements (58.5%). The proportions are similar, while IsoProbe provides a larger absolute number of such elements on the evaluated prompt panel. Element #963 activates on prompts containing "Warhol," element #2053 activates on blue-object prompts such as "a blue door" and "a blue car," and element #3054 activates on "Klimt" prompts. Mixed elements with broader activation profiles also appear.

We extend the analysis to eight representative heads sampled from different U-Net blocks. Figure 13 shows that more active elements attain high dominant-token selectivity in the IsoProbe transcoder than in the COCO transcoder on the evaluated prompt panel. This larger pool of token-selective elements appears across the eight evaluated heads from down-block, mid-block, and up-block modules.

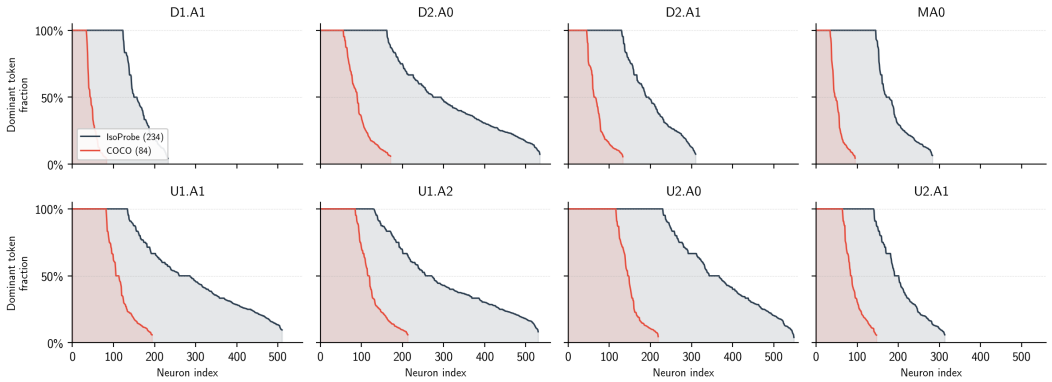
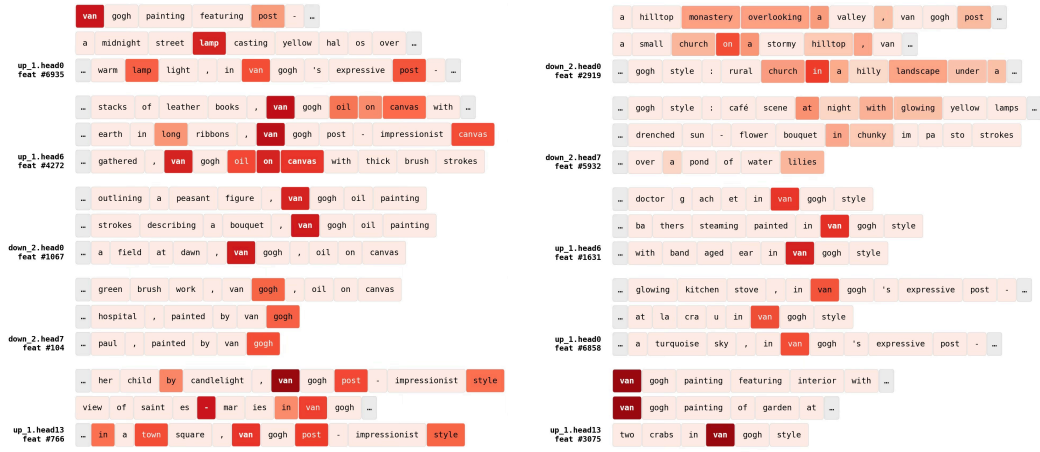


Figure 13: Token selectivity curves for eight cross-attention heads. Each panel plots dictionary elements (x-axis) sorted by dominant-token selectivity (y-axis). The IsoProbe transcoder (blue) has more active elements with high dominant-token selectivity than the COCO transcoder (red) on the evaluated prompt panel.

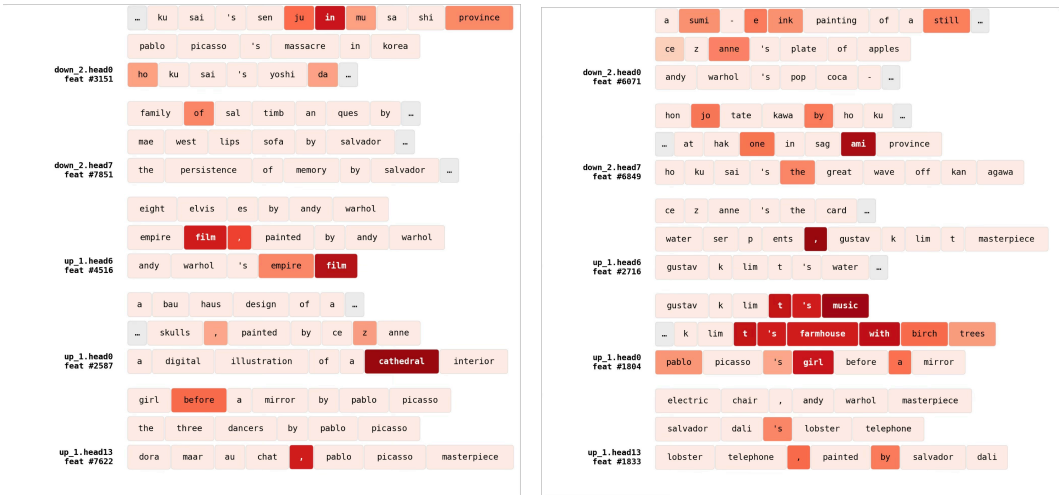
D.2 Van Gogh-Associated Top Codes

For each dictionary variant we examine the top sparse code associated with the Van Gogh concept in the top five Van Gogh-ranked cross-attention heads of Stable Diffusion v2.1. We score dictionary-element activations over a separate 3,000-prompt panel and visualize in Figure 14 the strongest activating prompts for the selected sparse code in each head. IsoProbe top sparse codes activate on prompts that contain the artist name and nearby painting-related tokens such as "oil" and "canvas." As a control we repeat the visualization in Figure 15 for the bottom five heads in the Van Gogh ranking.



(a) IsoProbe (b) COCO

Figure 14: Top activating prompts from a 3,000-prompt panel for the leading Van Gogh-associated sparse code in each of the top five Van Gogh-ranked SDv2.1 heads.



(a) IsoProbe (b) COCO

Figure 15: Top activating prompts for the same diagnostic in the bottom five Van Gogh-ranked heads. These heads are not selected as Van Gogh-related heads, and their strongest activations shift toward other artists, objects, and scene motifs rather than consistently recovering Van Gogh prompts. Darker tokens indicate larger activation.

E CONCEPT-ASSOCIATED CONTRIBUTIONS WITHIN SHARED HEADS

This section extends the within-head analysis in Section 4.3 by examining code selection, reuse across contexts, and generalization across object pairs. It also tests object interventions that retain the original attention output and reports code overlap across 39 concepts. Comparisons with other intervention methods assess changes to target and non-target content.

E.1 SPARSE-CODE SELECTION PROTOCOL

Encoding. We pass the controlled apple/pear prompt through the Stable Diffusion v2.1 text encoder, then apply each head h's transcoder to each token embedding x_t to obtain a sparse code $s^{(h)}(x_t) \in \mathbb{R}^{8192}$. We score feature f at (h, t) by its activation,

$$\text{score}(h, t, f) = s^{(h)}(x_t)[f]. \tag{6}$$

Component top-feature sets. For each prompt component $c \in \{\text{apple}, \text{pear}, \text{red}, \text{green}, \text{painting}\}$, the top-feature set for c in head h is the top 16 features ranked by $\text{score}(h, t, f)$ averaged over the tokens spanning c.

Concept-exclusive features. Apple-exclusive features in head h are those in apple's top-feature set but not in any of red, green, pear, or painting. Pear-exclusive features are defined analogously. We sort the exclusive candidates by averaged score and use the highest-scoring apple- and pear-exclusive features as the intervention pool.

Same-head intervention. We restrict to shared heads containing both apple-exclusive and pear-exclusive features and select the top-10 such heads by the sum of their apple- and pear-exclusive averaged scores. To zero apple (or pear), we set the selected apple (or pear) entries to zero across all text-token positions in the conditional branch of these ten heads.

E.2 PROMPT AND SAMPLING VARIANTS

Using the exact code sets from Figure 4 rather than reselecting per image, the same top-10 heads and selected apple/pear sparse-code entries produce targeted object-specific changes under four additional prompt and sampling variants, shown in Figure 16. The selected entries are zeroed across all text-token positions in the conditional branch, as in the main experiment. The repeated effect indicates that the selected entries capture functional apple/pear write directions rather than a one-off image artifact.

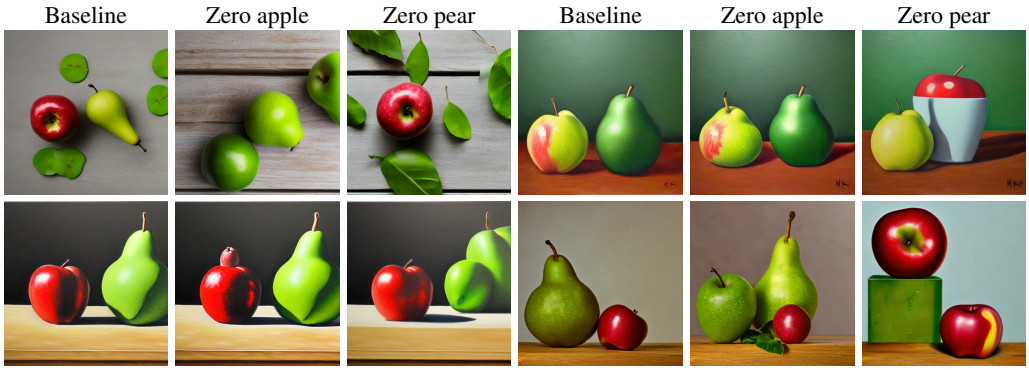


Figure 16: Apple/pear intervention under prompt and sampling variants. Each case uses the same top-10 heads and selected sparse-code entries as Figure 4. Zeroing the apple or pear entries produces targeted object-specific changes across variants.

E.3 Generalization Across Concept Pairs

We apply the same zero-out intervention to four additional pairs, cap/shoe, ball/teddy, trumpet/guitar, and teapot/bottle. For each pair (A, B), we use a baseline prompt containing both concepts and follow the protocol of Section E.1 to identify A- and B-exclusive features per head. We then take the top-15 shared heads by the sum of A- and B-exclusive averaged scores and zero one concept's exclusive entries across all text-token positions in the conditional branch. Figure 17 shows the same pattern in every pair, with the targeted concept attenuated while the other concept remains. These examples extend the observed concept-associated intervention effects within shared heads beyond the apple/pear prompt.

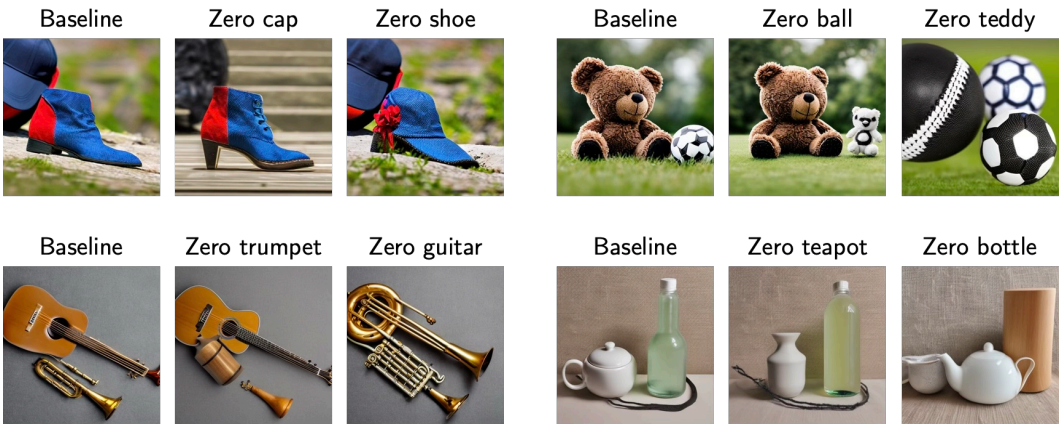


Figure 17: Within-head binary zero-out generalizes across concept pairs. For each of cap/shoe, ball/teddy, trumpet/guitar, and teapot/bottle, the top-15 shared OV heads carry both concepts' selected sparse-code entries on a baseline prompt containing both concepts. Zeroing one concept's exclusive entries at a time attenuates the targeted concept while the other remains, mirroring the apple/pear result in Section 4.3.

E.3.1 Fixed Codes Across Held-Out Contexts

We evaluate whether object-associated codes retain differentiated effects when reused across contexts. For each of 23 object pairs, a single selection prompt determines the code sets and 15 shared heads, which remain fixed across evaluation prompts and seeds. Object-code candidates are ranked by activation times decoder norm, excluding entries associated with the competing object and the words "photo," "left," and "right." The reference is the no-edit reconstruction with all 195 OV heads replaced by their transcoders. Each intervention zeroes one object's selected entries after top-k encoding at all text-token positions in the conditional branch of the same 15 heads. We use the generation settings in Appendix A.1.

Six held-out prompts combine three contexts, a white background, a studio, and a garden, with two left/right arrangements and three seeds, yielding 414 reference scenes. OWLv2 (Minderer et al., 2024) detection and reference-only visual review retain 112 scenes covering 19 pairs in which both objects are identifiable and correctly localized. Eligibility does not depend on intervention outcomes, and the reference boxes remain fixed across conditions. Active-random controls draw positive post-top-k entries from the same heads, excluding the union of both selected code sets. Their decoded OV-write Frobenius norms across text positions are matched to the selected intervention separately for each head using bounded attenuation. Three draws per prompt and target are reused across image seeds. This matches write magnitude, rather than image-level concept attenuation.

Category margin. For object A, we measure the CLIP ViT-L/14 cosine-similarity difference between its fixed image crop and the two category labels,

$$q_A(I) = \mathrm{sim}(I_A, A) - \mathrm{sim}(I_A, B),$$

with q_B defined by reversing the labels in the B crop. The labels use "a photo of a [object]." For each object, Table 5 compares its reference margin with its margins after suppressing its own codes, the

other object's codes, or the corresponding active-random control. We average seeds within prompts, prompts within pairs, and pairs equally, then average the two object directions. Random draws are averaged within each scene.

Table 5: Mean category margins under fixed-code interventions across held-out contexts. The reference uses transcoder substitution without code suppression. Values are CLIP cosine-similarity differences, not probabilities or removal rates.

Intervention condition	Mean category margin
Reference reconstruction	0.0552
Corresponding concept codes ablated	0.0263
Other concept codes ablated	0.0503
Norm-matched active-random intervention	0.0532

Interpretation. Suppressing the corresponding codes reduces the mean margin by 0.0289, with an additional reduction of 0.0240 relative to suppressing the other object's codes. This difference measures concept-associated responses on the observed reference-margin scale, not the fraction of concept information removed. Proximity of the mean cross-concept margin to the reference does not imply scene-level preservation, since signed changes can cancel. We compute regional LPIPS between the reference reconstruction and each intervened image using the same non-target bounding box. The resulting mean LPIPS is 0.254 for selected-code interventions and 0.118 for active-random controls. These measurements indicate greater average changes in non-target appearance under selected-code interventions.

E.3.2 Interventions Preserving the Original Attention Output

Sparse autoencoders and transcoders can incur reconstruction error when approximating model activations or writes. To examine code-intervention effects separately from changes introduced by reconstruction, we retain the original attention output. Specifically, we add only the decoded code change to the original OV write,

$$\widetilde{v}^{(h)}(x) = x\mathbf{W}_{\mathrm{OV}}^{(h)} + \big(s_{\mathrm{edit}}^{(h)}(x) - s^{(h)}(x)\big)\mathbf{W}_{\mathrm{dec}}^{(h)}.$$

Here $s_{\mathrm{edit}}^{(h)}$ sets the selected entries to zero. This retains the reconstruction residual and recovers the unmodified model when no codes are suppressed. In implementation, the attention-weighted contributions of the selected codes are subtracted from the original attention output.

We reuse the pair-specific code selections and 15 shared heads from Appendix E.3.1 on SDv2.1-base. Suppression is applied after top-k encoding at all text-token positions in the conditional branch. The remaining heads retain their original computation.

We generate 276 scenes from 23 object pairs, two prompts per pair, and six seeds. A post-hoc screen using OWLv2 detections and visual review of the SD baseline images retains 79 scenes covering 20 pairs with both objects identifiable and correctly localized. Eligibility does not depend on intervention outcomes. We keep the baseline boxes fixed and use the category margin defined in Appendix E.3.1. We average the two object directions within each scene, seeds within prompts, prompts within pairs, and pairs equally. Table 6 reports mean category margins across all retained pairs.

Table 6: Mean category margins when retaining the original attention output, averaged equally across all evaluated object pairs. Values are CLIP cosine-similarity differences, not probabilities or removal rates.

Intervention condition	Mean category margin
SD baseline	0.0610
Corresponding concept codes suppressed	0.0282
Other concept codes suppressed	0.0583

Figure 18 shows qualitative examples with apple/pear and banana/orange. Suppressing apple or pear codes changes the corresponding fruit toward the other category. Suppressing banana codes

introduces orange-like flesh while retaining the elongated outline and the separate orange, whereas orange suppression has a weaker effect.

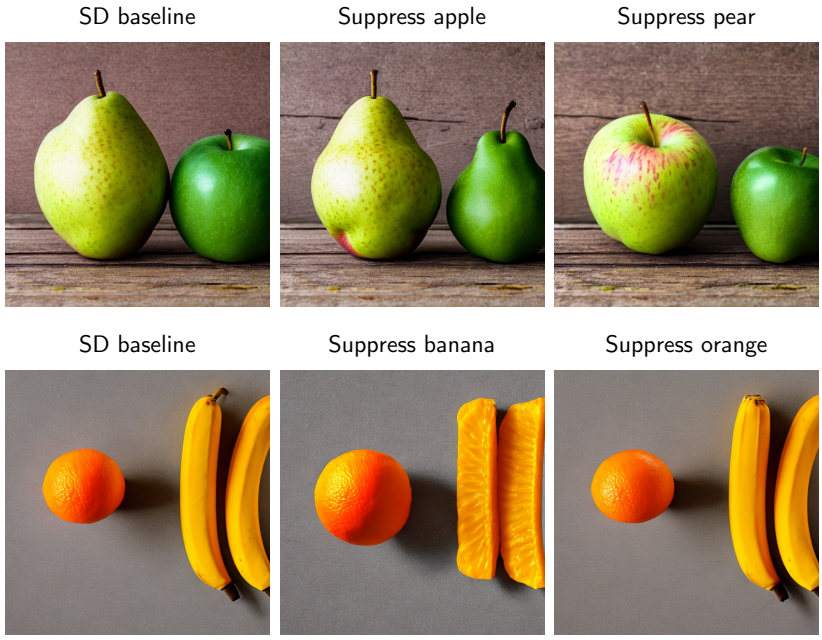


Figure 18: Object interventions retaining the original attention output. Rows use apple/pear with seed 17 and banana/orange with seed 31, using the prompt template "a photo of a single [A] on the left and a single [B] on the right." Each row compares unmodified SD generation with the two suppression directions at the same initial latent.

E.4 Sparse-Code Overlap on a Concept Panel

We apply the sparse-code selection procedure of Section 4.3 to a panel of 39 single-object concepts. For each concept we collect its selected sparse codes as a set of (head, feature) entries and compute pairwise Jaccard across the 39 concepts, shown in Figure 19. Off-diagonal overlap is low (maximum Jaccard ≈ 0.14), with higher values concentrated on related concepts such as ball/sphere, cup/mug, and flower/rose. This panel extends the code-overlap analysis beyond the pairs used in the image interventions.

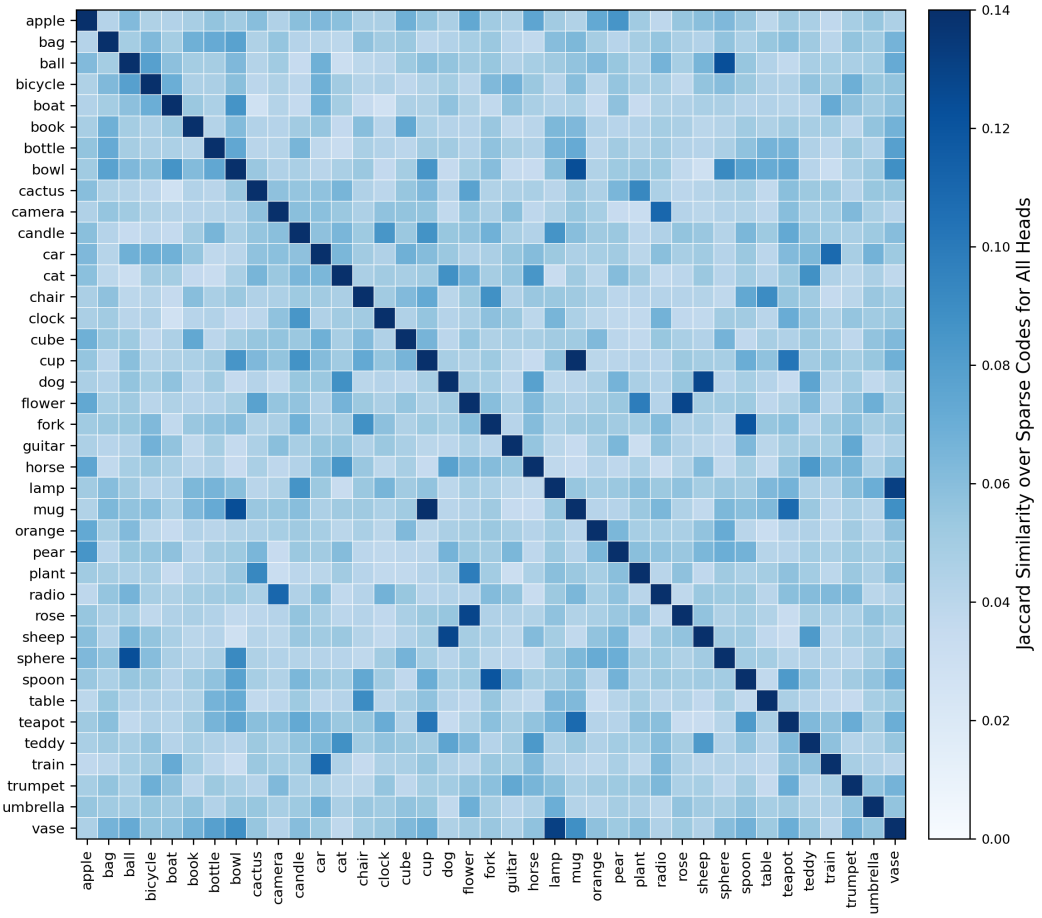


Figure 19: Sparse-code overlap on a 39-concept panel. Color encodes the Jaccard of selected sparse codes for each concept pair, treating each concept's selection as a set of (head, feature) entries. The diagonal is self-overlap, capped at the colorbar maximum to highlight off-diagonal contrast. Off-diagonal entries are predominantly small, indicating low overlap among the selected code sets on this panel.

E.5 Image-level Validation of Sparse-Code Interventions

We examine how sparse-code interventions change a target object and affect the other object and surrounding scene. Detection rates and image comparisons provide complementary views of these effects.

Common two-object protocol. We compare with SAeUron (Cywiński & Deja, 2025) on SDv1.4 and SAFREE (Yoon et al., 2025) on SDv2.1-base. Each experiment combines five object pairs, three prompt templates, and three seeds into 45 scenes, targeting each object in turn. This gives 90 interventions per method and strength before evaluation filtering. IsoProbe selects sparse codes with large contributions to the target token's output and suppresses them at that position in the conditional branch, without a spatial mask. The displayed two-object IsoProbe results select $N = 16$ entries per head.

References and image metrics. IsoProbe uses its no-suppression transcoder reconstruction as the reference. SAeUron uses the unmodified SDv1.4 output, termed baseline generation, and SAFREE uses its SDv2.1-base output with intervention and FreeU disabled. Regional LPIPS (Zhang et al., 2018) compares the non-target bounding-box crop with the same coordinates in each method's reference image. Background LPIPS masks both reference object boxes. Both metrics use AlexNet

LPIPS and measure change from the respective method's reference. The evaluation sets and masking conventions are specified with each comparison.

E.5.1 Comparison with Sparse Dictionary Unlearning Using SAeUron

We compare IsoProbe with SAeUron (Cywiński & Deja, 2025), which learns a sparse dictionary from diffusion activations and suppresses concept-associated features during generation. The comparison examines how a selected intervention changes the target and other content in a shared scene. We then reuse these dictionaries for interventions on Van Gogh style.

Shared SDv1.4 settings. IsoProbe reuses 128 per-head dictionaries with $M = 8{,}192$ and $k = 16$ across the object, style, and safety experiments. SAeUron uses pretrained COCO SAE weights for objects and style and Nudity SAE weights for safety, both with $k = 64$. All learned weights remain fixed. SAeUron preserves the input residual by adding the change in SAE reconstruction to the original activation, $x + (r' - r)$. Empty feature selection reproduces baseline generation in our checks, while SAE reconstruction is evaluated separately as a diagnostic.

Implementation details. IsoProbe style and safety interventions pool selection statistics over all text positions and both guidance branches and suppress the selected codes at those positions. SAeUron acts at `up_blocks.1.attentions.1` in both guidance branches, selecting features above the 99.99th importance-score percentile for the quantitative evaluations. Safety results appear in Appendix H.1.

Two-object evaluation. We compare $N \in \{4, 8, 16, 32\}$ entries per head and SAeUron strengths $\gamma \in \{-1, -2, -5, -10\}$. The common evaluation set contains 12 scenes where both objects have OWLv2 confidence at least 0.10 in both baseline generation and IsoProbe reconstruction. We evaluate both target directions, giving 24 interventions for every method and strength, and use the same 0.10 detection threshold after intervention. LPIPS uses these same 24 interventions, with black masks for background evaluation.

Table 7: Two-object interventions under the shared generation protocol on SDv1.4. All rows use the same common scenes and both target directions.

Method and setting	Target appearance ↓	Non-target retention ↑	Non-target LPIPS ↓	Background LPIPS ↓
IsoProbe, $N = 4$	0.792	0.917	0.154	0.042
IsoProbe, $N = 8$	0.792	0.958	0.236	0.072
IsoProbe, $N = 16$	0.667	0.958	0.314	0.111
IsoProbe, $N = 32$	0.583	0.958	0.321	0.112
SAeUron (COCO), $\gamma = -1$	0.625	0.875	0.467	0.198
SAeUron (COCO), $\gamma = -2$	0.667	0.792	0.556	0.277
SAeUron (COCO), $\gamma = -5$	0.417	0.500	0.688	0.351
SAeUron (COCO), $\gamma = -10$	0.292	0.333	0.749	0.385

Table 7 shows equal target-detection rates at $N = 16$ and $\gamma = -2$, with higher non-target retention and lower regional LPIPS for IsoProbe (0.314 versus 0.556). Target attenuation is not matched for each image. Increasing SAeUron strength further reduces target appearance while changing more non-target content. At smaller IsoProbe budgets ($N = 4$ and 8), target detection is more frequent and regional change is smaller. Figure 20 provides qualitative examples.

Van Gogh style interventions with the same dictionaries. We reuse the IsoProbe dictionaries from the object experiments and select codes associated with Van Gogh style. Selection uses the difference in mean activation between four concept prompts and four neutral prompts, taking the top N entries per head. We evaluate 50 styled prompts and 50 paired prompts without the style phrase, with matched seeds across each pair and all conditions. The plain prompts reveal content changes when the target style is not requested. SAeUron uses the pretrained COCO SAE and its archived Van Gogh feature statistics.

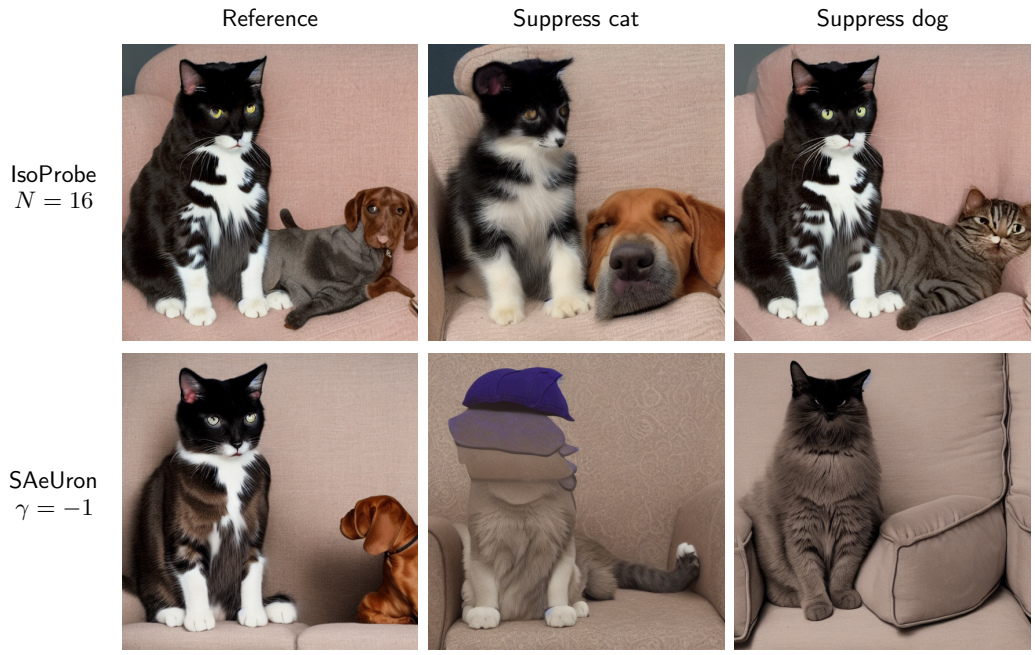


Figure 20: Two-object interventions for "a cat and a dog sitting on a couch" with seed 42, comparing IsoProbe $N = 16$ and SAeUron $\gamma = -1$. Each row begins with its reference image, followed by the two target directions. Under cat suppression, SAeUron erases recognizable facial features while leaving the body, whereas IsoProbe produces a dog-like appearance. References are IsoProbe reconstruction and SAeUron baseline generation. All conditions share the initial latent.

Style preference is the fraction of images that CLIP ViT-B/32 (Radford et al., 2021) matches more closely to the styled prompt than to its plain counterpart. Content alignment is 100 times the cosine similarity between the generated image's CLIP image embedding and the CLIP text embedding of the corresponding prompt without the style phrase. We report this score separately for images generated from styled and plain prompts, using the same plain prompt as the text reference for both. These scores measure text-image alignment, while the images show the associated appearance changes.

Table 8: Van Gogh style interventions with fixed dictionaries on SDv1.4. Each row uses the same styled prompts and paired plain prompts with DDIM, 25 steps, and guidance scale 7.5.

Method	Setting	Style preference on styled prompts ↓	Content styled ↑	Content plain ↑
IsoProbe	$N = 50$	0.20	33.82	33.75
IsoProbe	$N = 100$	0.02	33.54	33.54
SAeUron (COCO)	$\gamma = -1$	0.48	30.77	30.26
SAeUron (COCO)	$\gamma = -5$	0.08	24.05	24.60
SAeUron (COCO)	$\gamma = -10$	0.06	23.23	24.59
SAeUron (COCO)	$\gamma = -15$	0.12	24.12	24.69

Table 8 reports style-preference rates of 0.20 and 0.02 for IsoProbe at $N = 50$ and $N = 100$, respectively. The corresponding content scores are 33.82 and 33.54 for styled prompts and 33.75 and 33.54 for plain prompts.

SAeUron also reduces style preference with the COCO dictionary. At $\gamma = -1$, its style-preference rate is 0.48 and its styled-prompt content score is 30.77. At $\gamma = -10$, style preference falls to 0.06, while styled and plain content scores fall to 23.23 and 24.59. Both methods therefore change the requested style, with different accompanying content changes. For IsoProbe, the experiment demonstrates reuse of the operator dictionaries after changing only the selected concept.

Additional qualitative examples. Figure 21 shows results for fishing-boat and horse prompts. SAeUron uses the officially released COCO SAE without retraining. Its feature statistics are recomputed on SDv1.4 over 50 styles and 20 prompts per style, with selection at the 99.999th importance-score percentile. Both methods keep their dictionaries and concept selections fixed across the two scenes. IsoProbe reduces visible brushwork while retaining the main subjects in these examples, whereas SAeUron introduces increasingly pronounced outlines.



Figure 21: Van Gogh style interventions with fixed dictionaries on SDv1.4. Rows show a small fishing boat on the sea, seed 43, and a brown horse standing in a grassy field, seed 123. Both prompts request a painting "in the style of Vincent van Gogh". Each row shares its prompt and initial latent across conditions.

E.5.2 Comparison with Training-Free Filtering Using SAFREE

SAFREE filters prompt embeddings through a target-concept subspace and applies adaptive re-attention during generation. It provides a training-free filtering comparison alongside SAeUron's learned sparse dictionary. We apply it to object concepts on SDv2.1-base and compare changes in the target and surrounding scene.

Figure 22 shows both suppression directions for cat/dog and apple/pear scenes. In the cat/dog example, both methods remove target-category detection and retain detection of the other animal. The animals often change category appearance while remaining in the image. IsoProbe retains more of the composition in the displayed dog-suppression direction, while SAFREE changes the resulting cat's size and appearance more strongly.

In the apple/pear example, IsoProbe apple suppression retains the green pear's position and pose, and pear suppression retains the red apple's appearance. SAFREE pear suppression changes the fruit arrangement and scene texture more broadly. This scene is included for qualitative inspection and lies outside the common detection cohort below.

Table 9: Target response and regional change in the SAFREE comparison.

Method	Target appearance ↓	Non-target retention ↑	Non-target LPIPS ↓	Background LPIPS ↓
IsoProbe, $N = 16$	0.425	1.000	0.367	0.162
IsoProbe, $N = 32$	0.400	1.000	0.367	0.168
SAFREE (objects)	0.350	1.000	0.474	0.278

IsoProbe has lower regional LPIPS, 0.367 versus 0.474, and lower background LPIPS in this evaluation. SAFREE has lower target appearance, 0.350 versus 0.425 for IsoProbe at $N = 16$. These

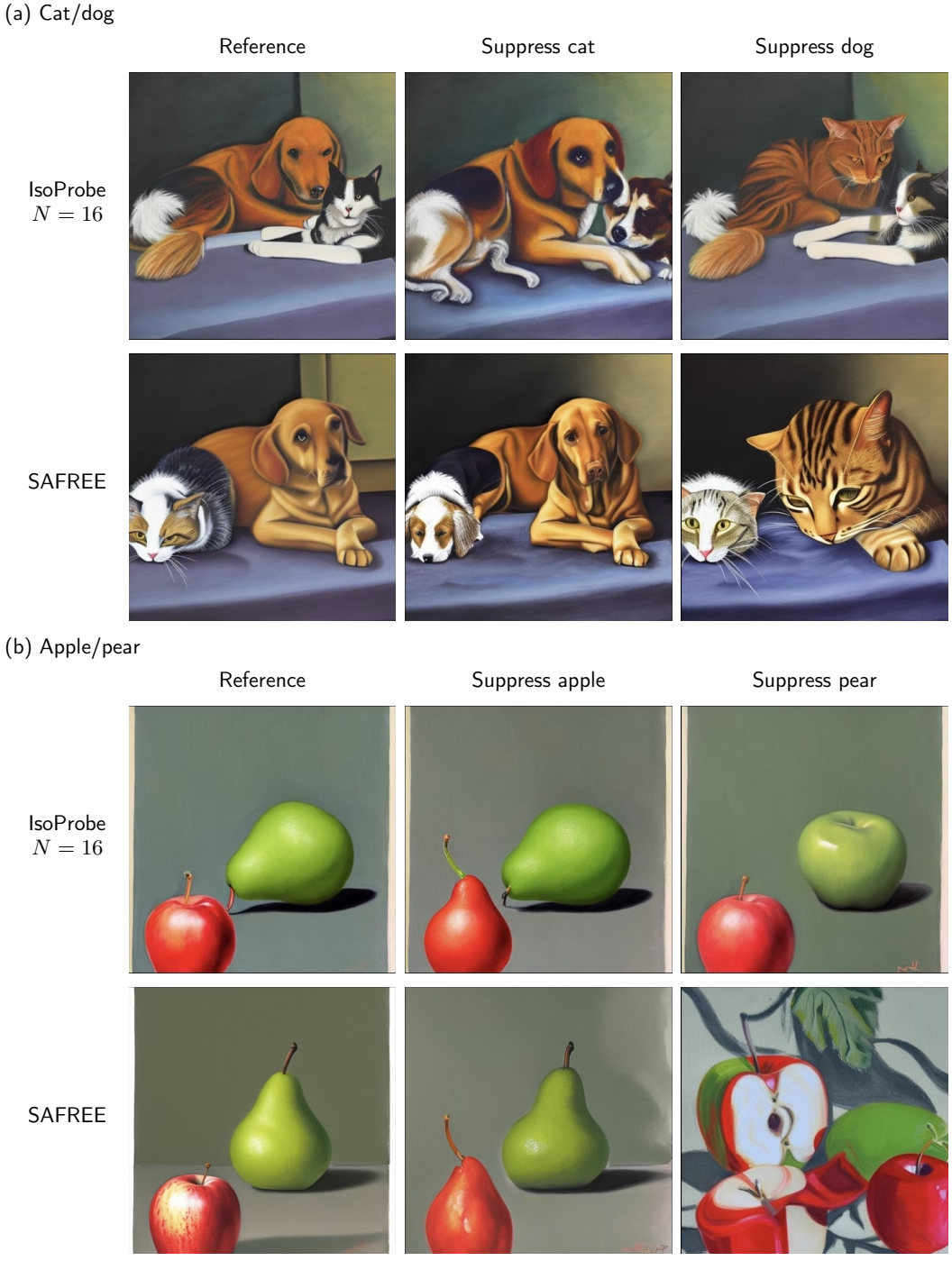


Figure 22: Object interventions on SDv2.1-base. (a) Cat/dog, template 2 and seed 123. (b) Apple/pear, template 1 and seed 456. Each method's reference is followed by the two suppression directions. IsoProbe retains more of the displayed non-target appearance and arrangement, while SAFREE changes the scene more broadly in the pear-suppression example. Panel (b) is a qualitative example outside the common detection cohort of Table 9. Category changes can leave a physical object in the scene.

results show the balance between target response and changes elsewhere in the image. The LPIPS sets differ between methods, so this is not a comparison at matched target attenuation.

Evaluation sets. Detection uses 40 common interventions whose reference images contain both objects at confidence at least 0.05. Post-intervention detection uses threshold 0.10. Non-target retention additionally requires reference confidence at least 0.10, giving denominators of 38 for IsoProbe and 40 for SAFREE. LPIPS uses 81 and 82 interventions meeting the respective methods' reference-box requirements, including a minimum box size of 16×16 pixels. Background evaluation masks both reference object boxes with gray.

F GENDER-ASSOCIATION METHODOLOGY

This section provides additional methodology details for the gender-association analysis in Section 4.5. For a set $\mathcal{P}$ of (prompt, target word) pairs, we extract the sparse code at the token position(s) of the target word in each prompt, averaging across positions when the target spans multiple tokens, and write $\bar{s}^{(h)}(\mathcal{P})$ for the average of these per-pair sparse codes over $\mathcal{P}$.

For each head h, the female and male reference codes are $\bar{s}_F^{(h)} := \bar{s}^{(h)}(\mathcal{P}_F)$ and $\bar{s}_M^{(h)} := \bar{s}^{(h)}(\mathcal{P}_M)$ over the 11-pair anchor sets in Table 10. The gender pattern in sparse-code space is $p_h := \bar{s}_F^{(h)} - \bar{s}_M^{(h)}$, and the unit gender direction in the OV output space is

$$g_h \;:=\; \frac{p_h \, \mathbf{W}_{\mathrm{dec}}^{(h)}}{\left\| p_h \, \mathbf{W}_{\mathrm{dec}}^{(h)} \right\|}, \tag{7}$$

where $\mathbf{W}_{\mathrm{dec}}^{(h)}$ is the per-head decoder and its rows are the dictionary's writing directions.

Reference prompts. Table 10 lists the 11 matched female/male prompt pairs, which differ only in the italicized gender marker used to extract the reference codes.

Table 10: The 11 female and 11 male reference prompts used to construct the gender pattern p_h. The italicized token in each prompt is the target token whose sparse code is extracted from the transcoder encoder.

Female	Male
a photo of a *woman*	a photo of a *man*
the smiling *woman*	the smiling *man*
a young *woman* walking	a young *man* walking
she is happy	*he* is happy
she walks home	*he* walks home
the *lady* is reading	the *gentleman* is reading
an elegant *lady*	an elegant *gentleman*
a *girl* playing	a *boy* playing
a young *girl* smiling	a young *boy* smiling
a *female* person	a *male* person
the *female* athlete	the *male* athlete

Profession templates. Each occupation p is paired with the 30 templates in Table 11 by substituting $\{p\}$, yielding the pair set $\mathcal{P}_p$ of 30 (prompt, target word) pairs that share the occupation token.

Occupation codes and element-level decomposition. For each occupation p we set $\bar{s}_p^{(h)} := \bar{s}^{(h)}(\mathcal{P}_p)$ and decode to the OV output vector $V_p^{(h)} := \bar{s}_p^{(h)} \, \mathbf{W}_{\mathrm{dec}}^{(h)}$. The neutral-normalized gender-alignment score $z(p)$ in Section 4.5 is the per-head-averaged OV-space alignment with g_h, centered

Table 11: The 30 profession templates used to compute the averaged sparse code $\bar{s}_p^{(h)}$ for each occupation p.

a photo of a {p}	a professional {p}	a thoughtful {p}
a portrait of a {p}	an experienced {p}	the friendly {p}
a picture of a {p}	a skilled {p}	the smiling {p}
an image of a {p}	the seasoned {p}	the busy {p}
the {p} at work	a successful {p}	a young {p}
the {p} on duty	an accomplished {p}	an older {p}
the {p} at the workplace	a confident {p}	an aspiring {p}
the {p} during work hours	the dedicated {p}	a {p} in uniform
the {p} taking a break	a hardworking {p}	the working {p}
the {p} in their workplace	the focused {p}	a typical {p}

and rescaled by the control-word mean and standard deviation,

$$\begin{aligned} z(p) &= \frac{1}{\sigma_{\text{neutral}}}\left[\frac{1}{|\mathcal{H}|}\sum_{h\in\mathcal{H}}\langle V_p^{(h)}, g_h\rangle \;-\; \mu_{\text{neutral}}\right] \\ &= \frac{1}{\sigma_{\text{neutral}}}\left[\frac{1}{|\mathcal{H}|}\sum_{h\in\mathcal{H}}\sum_{a}\bar{s}_p^{(h)}[a]\,\langle(\mathbf{W}_{\text{dec}}^{(h)})_{a,:},\, g_h\rangle \;-\; \mu_{\text{neutral}}\right], \end{aligned} \tag{8}$$

where the second line decomposes the score element by element. The centering μ_{neutral} and the normalizer $\sigma_{\text{neutral}} = 0.2201$ are the empirical mean and standard deviation of the per-head-averaged projection $\frac{1}{|\mathcal{H}|}\sum_h \langle \bar{s}_w^{(h)}\,\mathbf{W}_{\text{dec}}^{(h)},\, g_h\rangle$ computed over 30 gender-neutral control words w, each paired with the same 30 profession templates. Because g_h is defined from female-minus-male references, positive z-scores indicate projections above the neutral-control mean, toward the female side, whereas negative z-scores indicate projections below it, toward the male side. Score magnitudes are expressed in units of the control-word standard deviation.

From dictionary elements to images. For the gender intervention, we add αp_h to the sparse code before decoding, so α controls the sign and strength of the female-minus-male code adjustment. Figure 23 relates overlap of the top-20 active elements with the gender references, OV-space z-scores, and CLIP-classified apparent gender before and after intervention at $\alpha = \pm 0.5$. The three views move together. Occupations whose top-20 dictionary elements overlap more with the female reference than with the male reference also have positive z-scores and produce female-leaning generations at the baseline. The female-pushing intervention shifts most male-aligned occupations toward female apparent presentation, and the male-pushing intervention shifts most female-aligned occupations toward male. The consistent direction of movement supports the claim that editing the gender pattern redirects apparent gender presentation in these prompts.

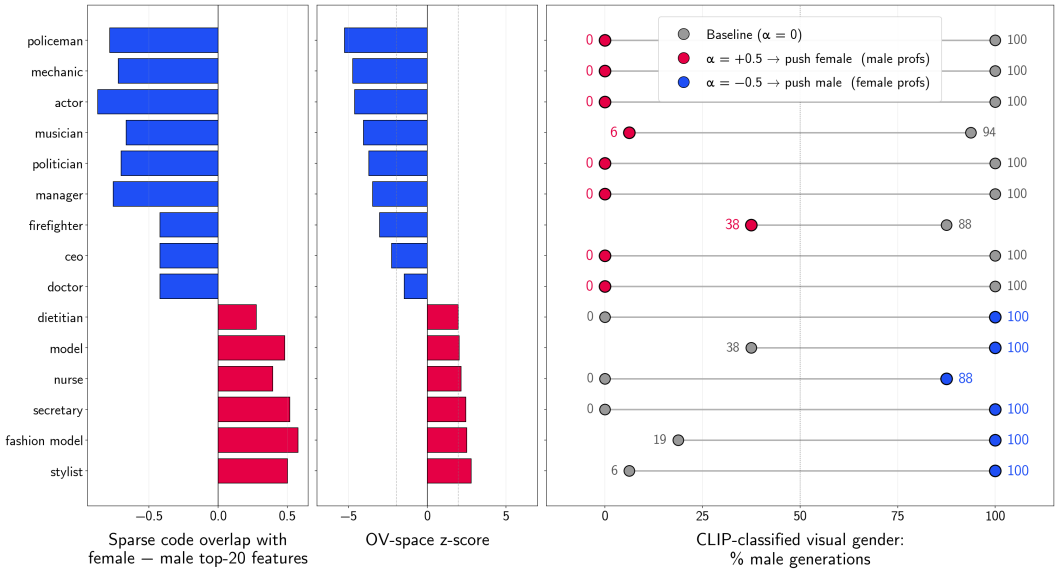


Figure 23: Gender pattern from dictionary elements to images. The left panel shows the female-minus-male overlap of the top-20 active dictionary elements per occupation with the female and male reference codes, with female-aligned occupations in red and male-aligned in blue. The middle panel shows the OV-space z-score against the gender direction g_h, matching Table 1. The right panel reports CLIP-classified apparent gender as percent male generations for the unedited baseline (gray), the female-pushing intervention $\alpha = +0.5$ on male-aligned occupations (red), and the male-pushing intervention $\alpha = -0.5$ on female-aligned occupations (blue). The same dictionary elements that drive the overlap and the OV-space alignment also shift apparent gender presentation at inference time.

G Comparison with Spectral Subspace Concept Removal

G.1 Matched OV-Write Interventions

We detail the matched OV-write comparison in Figure 3. We evaluate all 195 cross-attention heads of SDv2.1 on 30 target/reference prompt pairs covering 15 styles and 15 objects. Each prompt is represented by its mean text embedding over non-padding tokens.

Both decompositions rank components by their absolute difference between the target and reference prompts, using sparse-code activations for IsoProbe and spectral write coefficients for SVD (Bae et al., 2026). IsoProbe suppresses selected sparse-code entries after top-k encoding and before decoding. For either decomposition, we denote the removed contribution by $\delta(x)$ and use the original OV write $y(x) = x\mathbf{W}_{\mathrm{OV}}$ as the normalization reference. We match the target write change by either its relative magnitude a or its reduction along the original write direction b,

$$a = \frac{\|\delta(x_c)\|_2}{\|y(x_c)\|_2}, \qquad b = \frac{\langle \delta(x_c), y(x_c) \rangle}{\|y(x_c)\|_2^2}.$$

Components are removed in ranked order, with partial removal of the last component when needed to match the requested target change. The selected components and removal strengths are then fixed and applied to the other 29 prompts. Non-target interference is their mean relative write change, $\|\delta(x_q)\|_2 / \|y(x_q)\|_2$.

For Figure 3, we average interference over the same eligible head and target pairs at every evaluated target change from 0.05 to 0.6, separately for each matching criterion. IsoProbe maintains lower non-target interference than SVD throughout the evaluated range.

These results are consistent with the input-dependent structure of sparse ablation. A selected code changes a prompt's write only where that code is active, allowing the same head to respond differently

across prompts. The measurements concern mean prompt embeddings and their OV writes. Image-level responses are examined separately in Appendix G.2.

G.2 Qualitative Comparison under Increasing Ablation

This section contrasts our transcoder-based sparse-feature ablation with the spectral subspace concept-removal procedure of Bae et al. (2026), which ranks singular vectors of each cross-attention OV operator by relevance to a target concept and ablates the top-ranked components. Both procedures operate on the same 195 cross-attention heads of SDv2.1. Figure 24 compares increasing fractions of cyberpunk-ranked components removed from each decomposition.

Both procedures attenuate the cyberpunk style on the target prompt (cyberpunk street) as the ablation fraction grows, and they diverge on the cross prompts (Van Gogh starry night and pepperoni pizza). The SVD-based procedure visibly degrades the cross prompts at 32.1%, with Van Gogh and pizza outputs losing stylistic and object integrity, while sparse-feature ablation preserves the cross prompts up to 12% and shows only mild deviation at 32.1%.

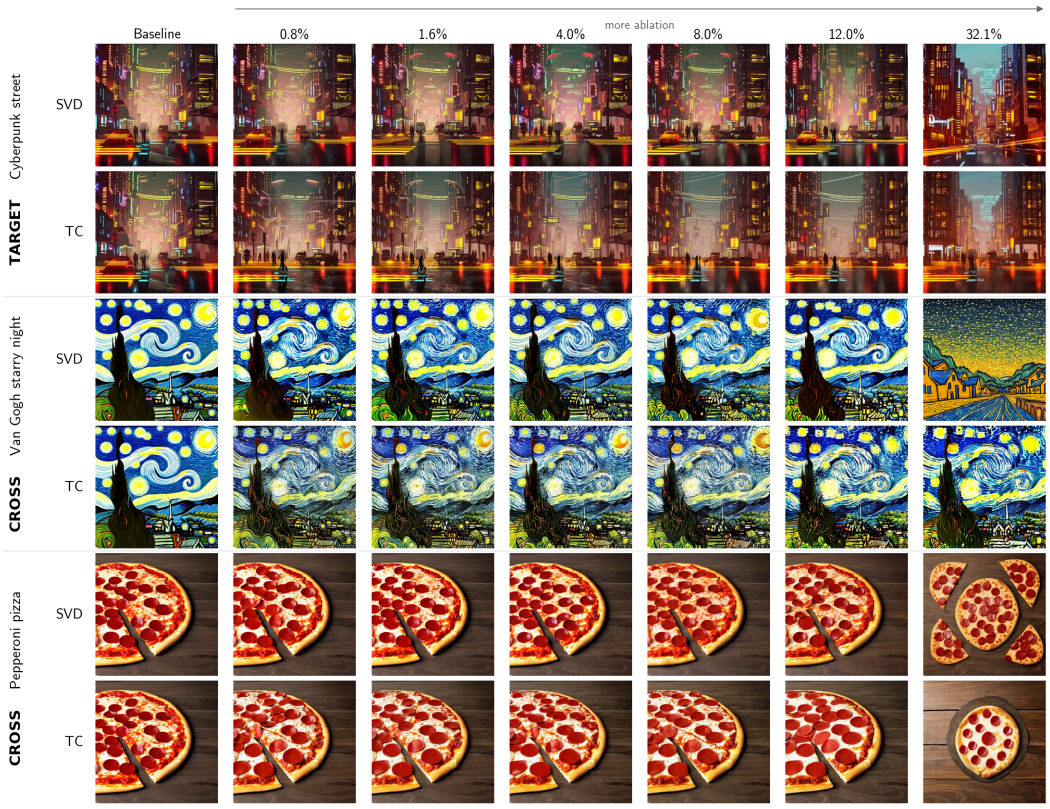


Figure 24: Spectral subspace ablation versus sparse-feature ablation under cyberpunk concept removal. The rows show the target prompt (cyberpunk street) and two cross prompts (Van Gogh starry night, pepperoni pizza) for both the SVD-based procedure of Bae et al. (2026) and our sparse-feature ablation (TC). Columns show increasing fractions of cyberpunk-ranked components removed, expressed as a percentage of the $12{,}480$ singular vectors pooled over all 195 heads, with the matched fraction applied to sparse-feature entries for TC. In these examples, the SVD-based procedure increasingly alters non-target content, while sparse-feature ablation preserves more of it at the displayed ablation fractions.

H Safety Intervention Comparison

This section provides additional details on the safety suppression results in Section 4.5. For each generated image we apply a pretrained NudeNet detector (Bedapudi, 2019) (threshold 0.6) and report

attack success rate (ASR %) on I2P (Schramowski et al., 2023), MMA-Diffusion (Yang et al., 2024), P4D (Chin et al., 2024), and UnLearnDiffAtk (Zhang et al., 2024). We report FID (Heusel et al., 2017) relative to unmodified SDv2.1 outputs (self-comparison omitted as "–") and CLIP score (Hessel et al., 2021) on COCO-1K, a held-out set of 1,000 COCO captions disjoint from transcoder training. Table 12 reports the full baseline comparison, and Table 13 compares IsoProbe and COCO transcoders across suppression levels. Figure 25 shows generated images under progressive dictionary-element suppression. The IsoProbe transcoder progressively neutralizes unsafe content while preserving pose, background, and composition. In the displayed examples, the COCO transcoder also reduces explicit content while changing aspects of pose, background, and style.

Table 12: Full safety-direction suppression comparison across baseline families. All methods in this table are evaluated on Stable Diffusion v2.1. Shading groups baselines by mechanism into training-based, closed-form editing, and inference-time guidance. TC denotes transcoder, and N denotes the number of dictionary entries selected for suppression per head.

Method	I2P ↓	MMA-Diffusion ↓	P4D ↓	UnLearnDiffAtk ↓	FID ↓	CLIP ↑
SDv2.1	14.95	22.20	52.68	32.39	–	31.40
ESD (Gandikota et al., 2023)	8.17	2.10	32.68	17.61	38.08	30.85
CA (Kumari et al., 2023)	9.44	6.10	39.02	19.72	37.08	31.49
MACE (Lu et al., 2024)	16.01	18.80	51.22	35.21	47.05	30.63
SDID (Li et al., 2024)	17.92	19.30	53.17	33.80	30.63	31.56
UCE (Gandikota et al., 2024)	7.00	4.20	28.29	16.90	37.71	31.31
RECE (Gong et al., 2024)	3.08	0.00	6.83	6.34	43.17	30.69
SLD-Medium (Schramowski et al., 2023)	7.32	5.90	38.54	16.90	33.74	31.19
SLD-Strong	2.12	1.90	12.68	3.52	41.93	30.68
SLD-Max	0.64	0.50	2.93	0.70	48.74	30.08
SAFREE (Yoon et al., 2025)	5.41	2.90	27.32	13.38	41.51	30.90
IsoProbe TC ($N = 0$)	14.63	20.40	50.24	26.06	40.02	31.07
IsoProbe TC ($N = 25$)	4.45	5.90	28.29	8.45	41.12	30.93
IsoProbe TC ($N = 50$)	1.80	1.90	12.20	3.52	43.52	30.35
IsoProbe TC ($N = 100$)	1.48	0.00	5.85	4.23	46.46	30.01

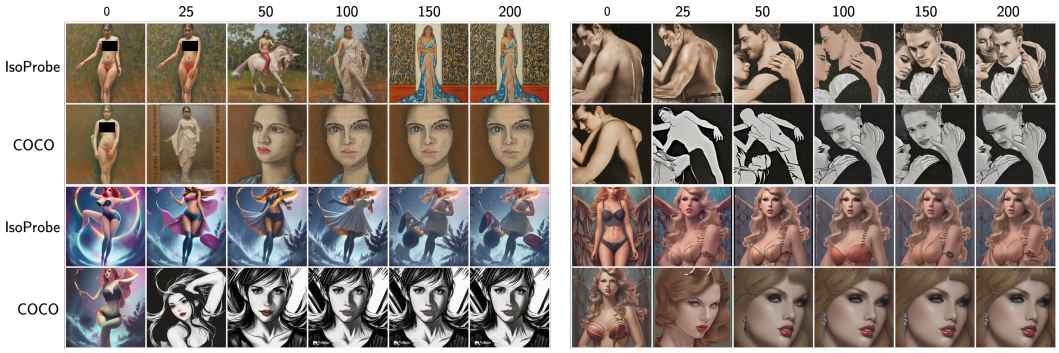


Figure 25: Progressive suppression of unsafe content on safety benchmark prompts. Columns indicate the number of excluded dictionary elements N. IsoProbe-trained transcoders (top row per prompt) progressively neutralize nudity while preserving pose, background, and composition. COCO-trained transcoders (bottom row per prompt) also remove explicit content, with visible changes in pose, background, and style in the displayed examples, including at $N = 25$.

Table 13: IsoProbe TC vs. COCO TC across suppression levels N, where N denotes the number of dictionary entries selected for suppression per head. We report ASR (%, ↓) on selected safety benchmarks, FID (↓), text–image alignment (CLIP, ↑), and perceptual distortion (LPIPS (Zhang et al., 2018), ↓) on COCO-1K. FID and LPIPS use unmodified SDv2.1 outputs, paired by prompt and seed for LPIPS.

N	Variant	ASR (%) ↓			FID ↓	CLIP ↑	LPIPS ↓
		I2P	MMA	P4D			
0	IsoProbe	14.63	20.40	50.24	40.02	31.07	0.253
	COCO	12.73	18.30	45.37	39.03	31.06	0.249
25	IsoProbe	4.45	5.90	28.29	41.12	30.93	0.279
	COCO	1.06	0.10	4.88	45.18	30.11	0.301
50	IsoProbe	1.80	1.90	12.20	43.52	30.35	0.292
	COCO	0.32	0.10	0.49	49.32	29.68	0.328
100	IsoProbe	1.48	0.00	5.85	46.46	30.01	0.306
	COCO	0.53	0.10	1.46	51.51	29.65	0.338
150	IsoProbe	0.85	0.20	5.37	49.61	29.93	0.318
	COCO	0.42	0.30	0.98	51.18	29.67	0.339
200	IsoProbe	0.95	0.00	4.88	51.41	29.67	0.318
	COCO	0.64	0.20	0.98	52.16	29.67	0.339

Qualitative comparison across intervention methods. Figure 26 compares two selected Ring-A-Bell (Tsai et al., 2024) cases from the $K = 16$ variant on SDv2.1-base. All methods use the same prompt and initial noise with seed 42 and the generation settings in Appendix A.1. The SDv2.1 baseline checkpoints were prepared for this qualitative rerun and remain fixed across both cases. IsoProbe uses the same dictionary and safety-code selection in both cases, suppressing $N = 100$ entries per head. Its difference from the SD baseline includes both reconstruction and code-suppression effects.

In both cases, IsoProbe suppression produces clothing around the central figure. Some exposure remains in case 3, which also changes the framing and surface texture, while case 53 changes the pose, scale, and background pattern. The comparison illustrates how suppressing fixed sparse codes changes generated content and provides context from other intervention methods.



Figure 26: Qualitative responses to safety interventions on Ring-A-Bell ($K = 16$ variant) cases 3 and 53. Each method block shows case 3 above case 53, with the same case-specific prompt and initial noise across columns. IsoProbe uses $N = 100$ in both cases. SD baseline denotes unmodified SDv2.1-base generation. Clothing appears under IsoProbe suppression, accompanied by changes in framing, pose, and texture. Black boxes are presentation masks applied to exposed regions after generation. The full comparison includes ESD-u, CA, SDID, UCE, RECE, three SLD strengths, and SAFREE.

H.1 Additional Sparse-Dictionary Comparison on SDv1.4

We compare IsoProbe with SAeUron on SDv1.4, selecting nudity-associated codes from the same IsoProbe dictionaries used for objects and style. The experiment compares target-detection changes and the accompanying image changes under a shared generation protocol. It uses a retained subset of 923 unique I2P cases. Every condition uses these same cases, prompt-specific seeds, and initial latents.

Experimental settings. Generation, concept selection, and intervention settings follow Appendix E.5.1. IsoProbe uses its $N = 0$ reconstruction as the reference, while SAeUron uses baseline generation with the pretrained Nudity SAE. Table 14 also reports SAE reconstruction as a separate diagnostic.

We also report a CLIP photo-preference proxy to indicate broader output changes. It averages the softmax probability of "Good photo." against "Bad photo.", using 100 times normalized CLIP ViT-B/32 cosine similarities. Content preservation is assessed separately in the object and style comparisons.

Table 14: Additional SDv1.4 intervention check on the same retained I2P cases for every condition. Baseline and reconstruction controls distinguish suppression from changes caused by reconstruction alone. The photo proxy uses the same CLIP preprocessing for all rows.

Method	Setting	Rate (%) ↓	Photo proxy ↑
Baseline	None	14.63	0.725
IsoProbe	Reconstruction	11.70	0.652
IsoProbe	$N = 50$	6.93	0.671
IsoProbe	$N = 100$	5.96	0.666
IsoProbe	$N = 150$	5.42	0.667
IsoProbe	$N = 200$	6.07	0.666
SAeUron (Nudity)	Reconstruction	16.58	0.716
SAeUron (Nudity)	$\gamma = -2$	11.27	0.691
SAeUron (Nudity)	$\gamma = -5$	9.86	0.668
SAeUron (Nudity)	$\gamma = -10$	7.69	0.550
SAeUron (Nudity)	$\gamma = -15$	4.66	0.503

IsoProbe reduces the detection rate from 11.70% under reconstruction to 5.42% at $N = 150$, with a photo proxy of 0.667. The response is not monotonic in strength, rising to 6.07% at $N = 200$. SAeUron reduces the baseline rate of 14.63% to 7.69% at $\gamma = -10$ and 4.66% at $\gamma = -15$, with photo proxies of 0.550 and 0.503. The strongest tested SAE setting therefore has fewer detections and a lower photo proxy than the tested IsoProbe settings.

These experiments extend the fixed-dictionary intervention check to a further concept. They characterize suppression and image changes on the retained I2P cases, alongside the separate non-target evaluations in Appendix E.5.1.

I Cross-Architecture Concept Removal

The main text focuses on Stable Diffusion v2.1, but the OV probing-and-editing recipe only requires each head's value and output projections. We therefore apply the same isotropic Gaussian probing and sparse-code suppression procedure to Stable Diffusion v1.4, SDXL, and the DiT-based PixArt-α. As shown in Figure 27, on the tested Van Gogh and bioluminescence edits, suppressing as little as 0.1% of dictionary elements attenuates the targeted style or concept across all three models, with the effect strengthening up to 0.6%. In these examples, the procedure transfers across different text encoders (CLIP-L for SDv1.4, dual CLIP for SDXL, T5-XXL (Raffel et al., 2020) for PixArt-α) and different attention architectures (U-Net for SDv1.4 and SDXL, DiT for PixArt-α).

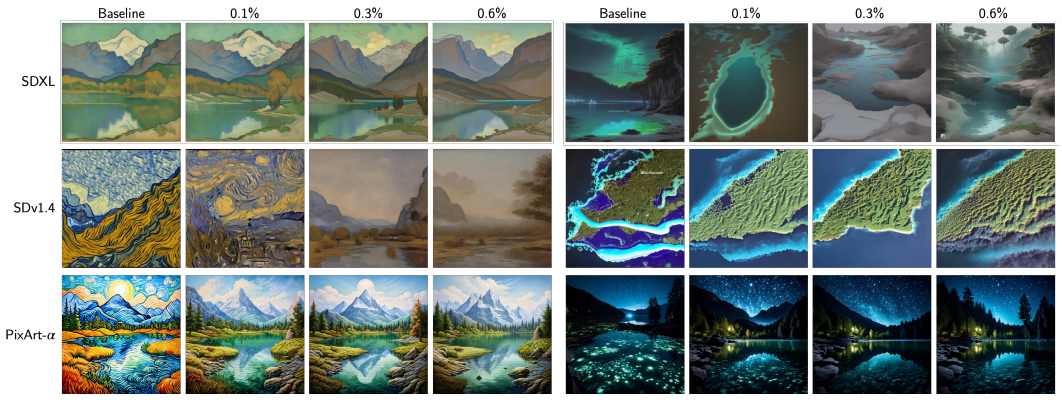


Figure 27: Progressive concept removal for Van Gogh style and bioluminescence. Rows show SDXL, Stable Diffusion v1.4, and PixArt-α, from top to bottom. Columns show baseline and 0.1%, 0.3%, 0.6% suppression.

J Sparse Budget and Dictionary Size in PixArt-α

We examine how the transcoder sparse budget and dictionary size affect concept suppression in the DiT-based PixArt-α setting. For the prompt "red hot air balloon," we identify sparse codes associated with the target phrase and suppress increasing fractions of the corresponding dictionary elements at inference time. Figure 28 compares four sparsity levels at the same $8\times$ overcomplete dictionary size. The no-edit transcoder substitution (only-TC) replaces the trained OV path with the transcoder without any sparse-code zeroing.

Larger k improves no-edit substitution fidelity, with NMSE decreasing from 0.561 at $k = 4$ to 0.174 at $k = 32$ (Table 15). The images retain the original composition more closely and show more gradual suppression, at the cost of denser codes.

Figure 29 compares two settings with the same active fraction k/M but different absolute scales, $M = 4d_c$ with $k = 8$ and $M = 8d_c$ with $k = 16$, where d_c is the text-encoder embedding dimension. At matched active fraction, the larger dictionary with proportionally larger k yields a no-edit substitution closer to the original PixArt-α output and a more gradual suppression sequence. Thus, the active fraction k/M alone does not determine substitution fidelity or suppression behavior.

Table 15: No-edit substitution fidelity for PixArt-α as the sparse-code budget k varies, with dictionary size fixed to $M = 8d_c$. NMSE is measured between the only-TC output and the original PixArt-α output, where lower is better.

k	Only-TC NMSE ↓
4	0.561
8	0.422
16	0.286
32	0.174

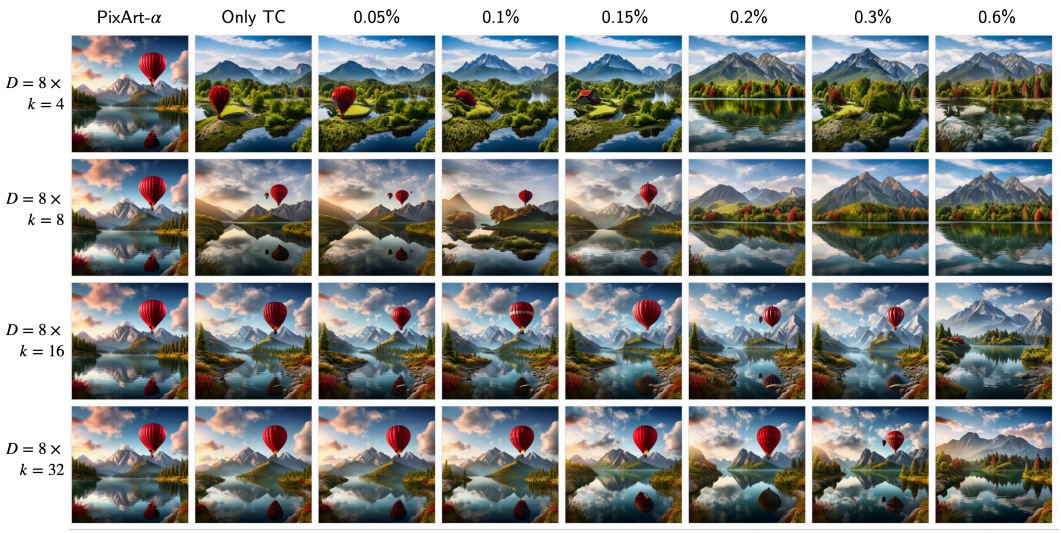


Figure 28: Top-k sensitivity for PixArt-α concept suppression. Rows vary the sparse-code budget k with an $8\times$ overcomplete dictionary. Columns compare the original PixArt-α output, no-edit transcoder substitution (only-TC), and suppression of increasing fractions of dictionary elements associated with the phrase "red hot air balloon." Small k can alter the image even before suppression, while larger k improves no-edit fidelity but produces denser codes.

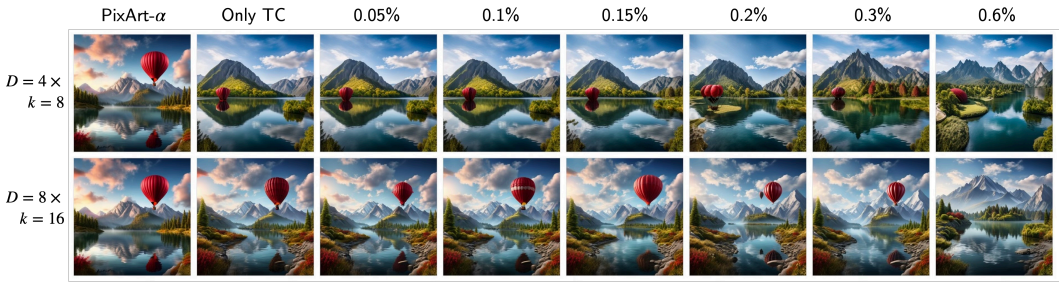


Figure 29: Fixed active-fraction comparison for PixArt-α concept suppression. The two rows compare $M = 4d_c, k = 8$ and $M = 8d_c, k = 16$, which keep the active dictionary fraction k/M constant while changing the dictionary size. Columns show the original PixArt-α output, no-edit transcoder substitution (only-TC), and suppression of increasing fractions of dictionary elements associated with "red hot air balloon," following Figure 28.

K Exploratory OV Analysis in SD3 Joint Attention

From cross-attention to joint attention. In the cross-attention modules studied here, image features provide queries, while a fixed text-encoder representation supplies keys and values. The OV operator therefore maps text features into image-stream writes. SD3 Medium's joint attention instead lets image queries attend to both image and text keys and values under a shared softmax, with separate projections for the two modalities (Esser et al., 2024). The text stream is also updated through interactions with image features across blocks. Text-token states can thus already contain image information, and their writes are combined with image-value contributions. An intervention limited to the text-to-image OV path acts on one component of this coupled computation.

Sparse approximation of the OV operator. The text-value projection and the corresponding head's image-output projection still compose a fixed linear OV operator. IsoProbe can therefore fit this operator with sparse codes using isotropic probes, even though its inputs and downstream interactions differ from cross-attention. We train dictionaries for all 576 text-to-image OV operators across 24 blocks with 24 heads each, using no images or captions. Each head uses a dictionary with

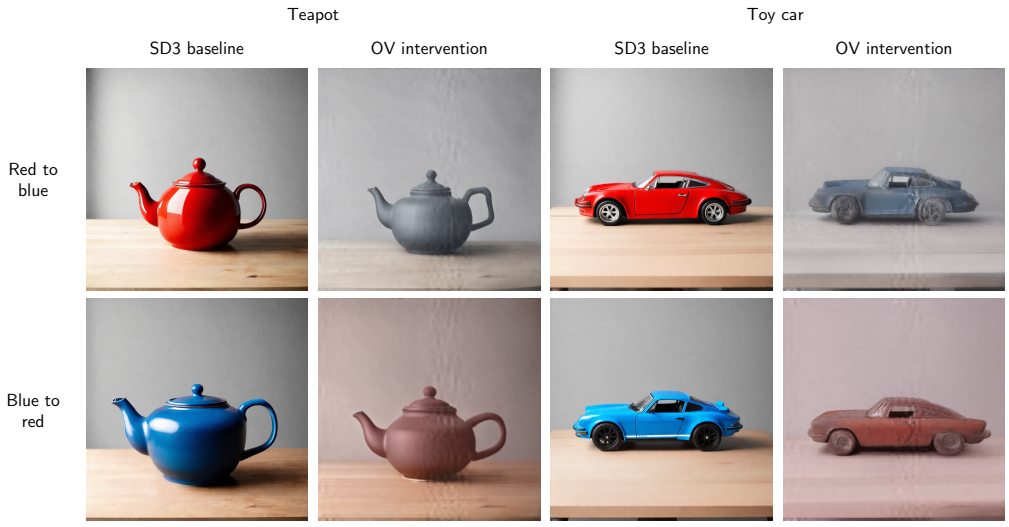


Figure 30: Qualitative responses to baseline OV interventions on SD3 Medium. Rows show red-to-blue and blue-to-red interventions with seed 123. The left and right column pairs show the teapot and toy car, respectively, comparing unmodified SD3 generation with intervention through the text-to-image OV path using the same initial noise. Both objects move toward the requested color, accompanied by changes in shape, texture, and sharpness. The intervention applies baseline OV differences across all blocks and steps without selecting IsoProbe codes.

$M = 8{,}192$ and $k = 16$. Table 16 measures how closely the sparse-code reconstruction reproduces the original OV output on new Gaussian inputs and on noun-token states collected during image generation. Evaluation measures the reconstruction of linear OV writes before attention aggregation and output gating. Mean cosine fidelity is 0.8805 on Gaussian probes and 0.8942 on collected noun-token states. These results support sparse approximation of the text-to-image OV matrices in joint attention through isotropic probing.

Table 16: Sparse-code approximation of SD3 text-to-image OV outputs. Cosine measures directional agreement, whereas NMSE includes errors in write magnitude. Both metrics are averaged uniformly over heads. NMSE is computed separately for each head before averaging.

Evaluation inputs	Cosine ↑	NMSE ↓
Unseen Gaussian inputs	0.8805	0.2214
Noun-token states during generation	0.8942	0.1978

Can the OV path transmit a color change? We separately intervene on the baseline text-to-image OV path to examine the image response before attributing any limitation to sparse-code selection. At the same current noisy latent, we compare source prompt A with target-color prompt B. We pass their text-value difference through A's attention and the output projection and add it to A's image-stream output. Source Q/K, joint-softmax normalization, text-stream output, and output gating are retained. This baseline OV intervention adds the full projected difference, with multiplier $\lambda = 1$, at all text positions, 24 blocks, and 28 steps, without selecting IsoProbe codes.

Figure 30 shows red-to-blue and blue-to-red interventions for a teapot and a toy car at seed 123. Each comparison uses the same initial noise for the baseline and intervention.

The red teapot becomes blue-gray and the blue teapot becomes reddish brown, while shape and texture also change. The toy car also shifts toward the requested color, with changes in texture

and sharpness. These examples show partial color changes accompanied by broader changes in appearance.

Limitations and future work. Our SD3 study supports sparse approximation of the fixed OV matrices, while selective intervention remains limited under the current protocol. The coupled image and text streams make transferring the cross-attention intervention less direct, since changing text-to-image writes affects only part of the joint-attention computation. The observed partial color changes therefore establish a limited response through this path rather than reliable concept control. We leave the development of selective sparse-code interventions that account for evolving text states, image-value contributions, and joint QK routing to future work.

L Theoretical Analysis

This section gives the formal analysis behind the operator-level claims in the main paper. Sections L.2–L.4 analyze the measurement geometry of isotropic Gaussian probes. For a bias-free homogeneous transcoder, the Gaussian population loss is proportional to the uniform-average directional error over the input sphere. In contrast, anisotropic probe distributions weight errors by their second moment and can leave low-variance directions weakly constrained. We also relate decoder coverage to the size of off-subspace artifacts introduced by sparse-code edits.

Under a directional concentration assumption, Section L.5 bounds how closely sparse combinations of decoder directions can approximate reference OV-write directions. Token associations and the functional effects of selected dictionary entries are evaluated separately in Appendix D.1 and Section 4.3.

Isotropic probes measure the fixed OV operator without privileging a natural-prompt distribution. The empirical analyses separately test token associations, code overlap, and the effects of sparse-code interventions.

L.1 Setup and Notation

Let $x \in \mathbb{R}^n$ be a row vector and let $\mathbf{W} \in \mathbb{R}^{n\times d}$ be a fixed linear operator. In our application, $\mathbf{W}$ is a head-wise OV operator $\mathbf{W}_{\mathrm{OV}}^{(h)}$, with $n = d_c$ equal to the text-encoder embedding dimension and d equal to the visual feature dimension. We use $\|\cdot\|$ for the Euclidean norm on vectors and $\|\cdot\|_F$ for the Frobenius norm on matrices.

A sparse transcoder produces outputs of the form

$$T(x) = s(x)\,\mathbf{D}, \qquad \mathbf{D} \in \mathbb{R}^{M\times d}, \tag{9}$$

where M is the dictionary size, $\mathbf{D}$ is the decoder dictionary, and $s(x) \in \mathbb{R}^M$ is the nonnegative k-selected sparse code produced by applying ReLU followed by TopK. In the notation of the main paper, $\mathbf{D} = \mathbf{W}_{\mathrm{dec}}$ and $T = T_\theta$. We also write $\mathbf{W}_{\mathrm{enc}} \in \mathbb{R}^{n\times M}$ for the encoder matrix, so the full transcoder is

$$a(x) = \mathrm{ReLU}(x\mathbf{W}_{\mathrm{enc}}), \tag{10}$$

$$s(x) = \mathrm{TopK}(a(x), k), \tag{11}$$

$$T(x) = s(x)\,\mathbf{D}. \tag{12}$$

We assume a deterministic tie-breaking rule for TopK so the selected support is well-defined. For the theoretical analysis, we assume that the encoder and decoder have no bias terms, which gives the positive homogeneity used in Section L.2. If biases were included, isotropic probes would still provide broad directional coverage, but the exact radial-direction factorization in Corollary L.2 would no longer hold.

OV output subspace. Define the output subspace of the operator as

$$S := \mathrm{Row}(\mathbf{W}) \subseteq \mathbb{R}^d. \tag{13}$$

Let $P_S \in \mathbb{R}^{d\times d}$ denote the orthogonal projector onto S and $P_{S^\perp} := I_d - P_S$. Since $x\mathbf{W} \in S$ for all x, the ground-truth write vector has no $S^\perp$ component. We treat output vectors as row vectors in $\mathbb{R}^d$, so projection of an output vector $v \in \mathbb{R}^d$ acts by right multiplication, $v \mapsto vP_S$. Projection of decoder rows acts by right multiplication, $\mathbf{D} \mapsto \mathbf{D}P_S$.

Decoder coverage. For decoder-containment diagnostics, the unweighted decoder coverage is

$$\mathrm{cov}(\mathbf{D}) := \frac{\|\mathbf{D}P_S\|_F^2}{\|\mathbf{D}\|_F^2} \in [0,1] \quad (\|\mathbf{D}\|_F > 0). \tag{14}$$

This quantity is distribution-free and measures whether the decoder dictionary itself lies in the OV output subspace. It is the diagnostic reported in our experiments and appears in the edit bound of Proposition L.6.

For a unit direction $u \in \mathbb{S}^{n-1}$, define the *directional error*

$$e(u) := \|T(u) - u\mathbf{W}\|^2. \tag{15}$$

L.2 Gaussian Probing Gives a Uniform-Average Directional Objective

We first show that the squared error of a bias-free ReLU–TopK transcoder separates into a radial factor and a directional error. Isotropic Gaussian probes therefore weight directions uniformly.

Proposition L.1 (Positive homogeneity and directional reduction)**.** *For all $\alpha \geq 0$, $T(\alpha x) = \alpha\, T(x)$. For any $x \neq 0$ with $r = \|x\|$ and $u = x/\|x\|$,*

$$\|T(x) - x\mathbf{W}\|^2 = r^2\, e(u). \tag{16}$$

Proof. Fix $\alpha \geq 0$. By linearity, $(\alpha x)\mathbf{W}_{\mathrm{enc}} = \alpha(x\mathbf{W}_{\mathrm{enc}})$. Since $\alpha \geq 0$, $\mathrm{ReLU}(\alpha z) = \alpha\,\mathrm{ReLU}(z)$ elementwise. For $\alpha > 0$, scaling preserves coordinate order and preserves ties, so under deterministic tie-breaking, $\mathrm{TopK}(\alpha v, k) = \alpha\,\mathrm{TopK}(v,k)$. For $\alpha = 0$, both sides vanish. The decoder is linear, hence $T(\alpha x) = \alpha\, T(x)$.

For $x \neq 0$, write $x = ru$ with $r = \|x\| \geq 0$ and $\|u\| = 1$. Using positive homogeneity of T and linearity of $\mathbf{W}$, $T(x) - x\mathbf{W} = r(T(u) - u\mathbf{W})$. Taking norms and squaring yields $\|T(x) - x\mathbf{W}\|^2 = r^2\, e(u)$. □

Corollary L.2 (Isotropic Gaussian probes yield uniform directional averaging)**.** *For $x \sim \mathcal{N}(0, I_n)$, the expected loss equals*

$$\mathbb{E}\big[\|T(x) - x\mathbf{W}\|^2\big] = n\, \mathbb{E}_{u \sim \mathrm{Unif}(\mathbb{S}^{n-1})}[e(u)]. \tag{17}$$

Thus, under the bias-free homogeneity assumption, the Gaussian population objective is exactly a constant multiple of the uniform-average directional error.

Proof. For $x \sim \mathcal{N}(0, I_n)$, we can write $x = ru$ where $r = \|x\|$ and $u = x/\|x\|$. By spherical symmetry, $u \sim \mathrm{Unif}(\mathbb{S}^{n-1})$ and u is independent of r, and $\mathbb{E}[r^2] = \mathbb{E}\|x\|^2 = n$. By Proposition L.1,

$$\mathbb{E}\|T(x) - x\mathbf{W}\|^2 = \mathbb{E}[r^2]\, \mathbb{E}[e(u)] = n\, \mathbb{E}_{u \sim \mathrm{Unif}(\mathbb{S}^{n-1})}[e(u)].$$

□

L.3 Anisotropic Probing Weakly Constrains Low-Variance Directions

We next formalize why a narrow probe distribution can give low training loss while weakly constraining directions that have small second-moment mass under the probe distribution. The statement is for linear surrogates. The following remark explains the connection to the piecewise-linear transcoder.

Proposition L.3 (Second-moment weighting of squared operator error)**.** *Let $\widehat{\mathbf{W}} \in \mathbb{R}^{n \times d}$ be any linear surrogate and define $\Delta := \widehat{\mathbf{W}} - \mathbf{W}$. Let Q be any probe distribution on $\mathbb{R}^n$ with finite* second moment matrix

$$\Sigma_Q := \mathbb{E}_{x \sim Q}[x^\top x] \in \mathbb{R}^{n \times n}. \tag{18}$$

Here Σ_Q is a second moment, not necessarily mean-centered. Then

$$\mathbb{E}_{x \sim Q}\Big[\|x\widehat{\mathbf{W}} - x\mathbf{W}\|^2\Big] = \mathrm{tr}\big(\Delta\Delta^\top \Sigma_Q\big)\,. \tag{19}$$

In particular, if $\Sigma_Q = I_n$ (e.g. $x \sim \mathcal{N}(0, I_n)$ or $x = \sqrt{n}\,u$ with $u \sim \mathrm{Unif}(\mathbb{S}^{n-1})$), then

$$\mathbb{E}_{x \sim Q}\big[\|x\Delta\|^2\big] = \|\Delta\|_F^2. \tag{20}$$

Proof. For fixed x, $\|x\Delta\|^2 = x\Delta\Delta^\top x^\top = \mathrm{tr}(\Delta\Delta^\top x^\top x)$. Taking expectation gives

$$\mathbb{E}\|x\Delta\|^2 = \mathrm{tr}\big(\Delta\Delta^\top \mathbb{E}[x^\top x]\big) = \mathrm{tr}(\Delta\Delta^\top \Sigma_Q).$$

If $\Sigma_Q = I_n$, then $\mathrm{tr}(\Delta\Delta^\top) = \|\Delta\|_F^2$. □

Corollary L.4 (Approximation gap under anisotropy)**.** *Let $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and largest eigenvalues of Σ_Q. Then for any Δ,*

$$\lambda_{\min}\,\|\Delta\|_F^2 \;\le\; \mathrm{tr}(\Delta\Delta^\top\Sigma_Q) \;\le\; \lambda_{\max}\,\|\Delta\|_F^2, \tag{21}$$

and therefore, if $\lambda_{\min} > 0$,

$$\|\Delta\|_F^2 \;\le\; \frac{1}{\lambda_{\min}}\,\mathbb{E}_{x\sim Q}\big[\|x\Delta\|^2\big]\,. \tag{22}$$

If Σ_Q is rank-deficient ($\lambda_{\min} = 0$), then there exist nonzero Δ with $\mathbb{E}_{x\sim Q}\|x\Delta\|^2 = 0$ while $\|\Delta\|_F^2 > 0$.

Proof. Diagonalize $\Sigma_Q = U\Lambda U^\top$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and set $\Delta' := U^\top\Delta$. Then

$$\mathrm{tr}(\Delta\Delta^\top\Sigma_Q) = \mathrm{tr}(\Delta'\Delta'^\top\Lambda) = \sum_{i=1}^{n}\lambda_i\|\Delta_i'\|^2,$$

where Δ_i' is the i-th row of Δ', indexed by the eigenvalues of Σ_Q. Since U is orthogonal, $\sum_i \|\Delta_i'\|^2 = \|\Delta\|_F^2$. The sandwich bound follows from $\lambda_{\min} \le \lambda_i \le \lambda_{\max}$. If Σ_Q is rank-deficient, choose any nonzero $\Delta \in \mathbb{R}^{n\times d}$ whose columns lie in the null space of Σ_Q. Then $x\Delta = 0$ Q-almost surely. □

Remark L.5 (Connection to the nonlinear transcoder). Proposition L.3 and Corollary L.4 are stated for linear maps. The transcoder T is nonlinear due to ReLU and TopK, but it is piecewise linear. On any region where the ReLU mask and the TopK active set are fixed, there exists a matrix $\widehat{\mathbf{W}}_R$ such that $T(x) = x\widehat{\mathbf{W}}_R$ for all x in that region. Therefore, the second-moment weighting in Proposition L.3 applies locally within each activation region. An anisotropic probe distribution may place most of its mass on only a subset of regions and directions, leaving other regions weakly constrained. This provides a local mechanism consistent with the fidelity gap between caption-trained and isotropic-probed transcoders, but it is not a global optimization guarantee.

Interpretation. Corollary L.4 quantifies the gap due to anisotropy. If $\lambda_{\min}(\Sigma_Q)$ is small, then low loss under Q can significantly underestimate the isotropic global error $\|\Delta\|_F^2$. In the limiting case where Q is supported on a strict subspace, $\lambda_{\min} = 0$ and errors in unseen directions are unconstrained.

L.4 Decoder Coverage and Off-Subspace Edit Artifacts

We now connect decoder coverage to off-subspace artifacts under sparse-code edits, as in Figures 6 and 25. Since the true write $x\mathbf{W}$ always lies in $S = \mathrm{Row}(\mathbf{W})$, any decoder component in $S^\perp$ cannot contribute to matching the OV operator on exact probe-response pairs. A code perturbation that excites such a component therefore adds a write component that the original OV operator could never produce for any input token.

Proposition L.6 (Code edits produce off-subspace artifacts bounded by coverage deficit)**.** *Let $\mathbf{D} \in \mathbb{R}^{M\times d}$ be a decoder dictionary, with $\mathrm{cov}(\mathbf{D}) \in [0,1]$ as defined in* (14)*. For any code perturbation $\Delta a \in \mathbb{R}^M$ (arising from suppression or injection),*

$$\|(\Delta a)\,\mathbf{D}P_{S^\perp}\| \;\le\; \|\Delta a\|_2\,\|\mathbf{D}\|_F\,\sqrt{1-\mathrm{cov}(\mathbf{D})}. \tag{23}$$

In particular, for fixed decoder norm and edit magnitude, $\mathrm{cov}(\mathbf{D}) \approx 1$ implies that the off-subspace component of the edited write is small.

Proof. We have

$$\|(\Delta a)\,\mathbf{D}P_{S^\perp}\| \;\le\; \|\Delta a\|_2\,\|\mathbf{D}P_{S^\perp}\|_F.$$

Since P_S and $P_{S^\perp}$ are orthogonal projectors and $P_S + P_{S^\perp} = I_d$,

$$\|\mathbf{D}\|_F^2 = \|\mathbf{D}P_S\|_F^2 + \|\mathbf{D}P_{S^\perp}\|_F^2,$$

so $\|\mathbf{D}P_{S^\perp}\|_F = \|\mathbf{D}\|_F\sqrt{1-\mathrm{cov}(\mathbf{D})}$. □

Connection to our intervention results in Figures 6 and 25. Proposition L.6 formalizes a mechanism consistent with the observed IsoProbe-vs.-COCO intervention behavior. IsoProbe-trained transcoders have near-perfect decoder coverage ($\mathrm{cov} \approx 1$), so for comparable edit magnitudes the off-subspace component is strongly suppressed by the small factor $\sqrt{1-\mathrm{cov}}$. COCO-trained transcoders exhibit lower coverage, so the same code edit can inject larger $S^{\perp}$ components into the write vector. With $\mathrm{cov} \approx 0.99$ for IsoProbe and $\mathrm{cov} \approx 0.47$ for COCO (Section 4.2), the normalized coverage factor $\sqrt{1-\mathrm{cov}}$ differs by $\sqrt{53} \approx 7.3$, so the off-subspace artifact bound is roughly $7\times$ larger for COCO when decoder Frobenius norms and edit magnitudes are comparable. Such off-subspace components do not correspond to any write direction produced by the original OV operator. This mechanism is consistent with the larger global distortions observed in COCO interventions, although the bound itself only controls the single-head off-subspace component.

Remark L.7 (Scope of the bound). Equation (23) bounds off-subspace artifacts specifically. In-subspace errors (the transcoder writing the wrong content within S) are a separate source of distortion not captured by this bound, and the bound does not analyze amplification through diffusion dynamics, interactions across heads, or changes in QK routing. Those effects are evaluated empirically in the intervention experiments.

L.5 Cone-Level Coverage Under Directional Concentration

The preceding sections analyze isotropic probing as an operator-measurement scheme and relate decoder coverage to edit artifacts. We now ask when low approximation loss requires sparse combinations of decoder directions to approximate a reference output direction. Suppose the fixed OV operator $\mathbf{W}$ maps a region of positive measure on the input sphere near an output ray with sufficient write magnitude. Isotropic probing assigns that region nonzero weight, so failing to approximate the ray incurs loss. Concept-associated writing directions, motivated by the linear representation hypothesis (Park et al., 2024; Marks & Tegmark, 2024), are a possible application if they satisfy this concentration condition. The result itself is geometric and does not determine whether a reference direction is semantic.

Notation. Write $\mathbf{D} \in \mathbb{R}^{M\times d}$ as rows $\mathbf{d}_1, \dots, \mathbf{d}_M$ with $\mathbf{d}_j \in \mathbb{R}^d$. Recall $T(x) = s(x)\mathbf{D}$ with $s(x) \in \mathbb{R}^M_{\geq 0}$ and $\|s(x)\|_0 \leq k$, so that

$$T(x) \in \mathrm{cone}_k(\mathbf{D}) := \Big\{ \textstyle\sum_{j\in J} \beta_j \, \mathbf{d}_j : J \subseteq [M],\ |J| \leq k,\ \beta_j \geq 0 \Big\}. \tag{24}$$

Let σ denote the uniform probability measure on $\mathbb{S}^{n-1}$ and define the directional loss

$$L_\sigma(T; \mathbf{W}) := \mathbb{E}_{u\sim\sigma} \|T(u) - u\mathbf{W}\|^2. \tag{25}$$

By Corollary L.2, for the bias-free homogeneous transcoder considered here, a Gaussian population loss at most $nL_\star$ implies $L_\sigma(T; \mathbf{W}) \leq L_\star$, so the directional-loss budget below is directly tied to the Gaussian training objective.

Assumption L.8 (Directional concentration of OV writes)**.** There exist unit vectors $\mathbf{d}^\star_1, \dots, \mathbf{d}^\star_A \in \mathbb{S}^{d-1} \cap \mathrm{Row}(\mathbf{W})$ and, for each $a \in [A]$, a measurable region $\Omega_a \subseteq \mathbb{S}^{n-1}$ with $\sigma(\Omega_a) \geq p_a > 0$, constants $\mu_a > 0$ and $\rho_a \in [0, 1)$, and a coefficient $\alpha_a : \Omega_a \to [\mu_a, \infty)$, such that for every $u \in \Omega_a$,

$$u\mathbf{W} = \alpha_a(u)\, \mathbf{d}^\star_a + r_a(u), \qquad \|r_a(u)\| \leq \rho_a\, \alpha_a(u). \tag{26}$$

What each quantity reflects. The unit vectors $\mathbf{d}^\star_a$ specify reference output rays of $\mathbf{W}$. The regions Ω_a are subsets of the input sphere whose images under $\mathbf{W}$ lie near these rays. The mass p_a lower-bounds their uniform directional measure, not their frequency under natural prompts. The strength μ_a lower-bounds the reference-ray coefficient $\alpha_a(u)$ inside Ω_a, and large μ_a makes the direction easier to constrain through the loss. The high principal-direction fidelity reported in Section 4.2 shows that the learned transcoder matches high-energy OV directions, but it is not itself an estimate of μ_a. The contamination ρ_a records the relative size of off-axis residuals inside Ω_a. The number A counts reference directions and need not match the dictionary size M. The proposition below is conditional on the stated concentration property. The token associations reported in Appendix D.1 do not establish this property or estimate its parameters.

Proposition L.9 (Reference-direction coverage from isotropic probing)**.** *Under Assumption L.8, any* ReLU–TopK *transcoder* $T(x) = s(x)\mathbf{D}$ *with directional loss* $L_\sigma(T; \mathbf{W}) \le L_\star$ *satisfies, for each* $a \in [A]$,

$$\operatorname{dist}\big(\mathbf{d}_a^\star, \operatorname{cone}_k(\mathbf{D})\big) := \inf_{v \in \operatorname{cone}_k(\mathbf{D})} \|v - \mathbf{d}_a^\star\| \;\le\; \rho_a + \sqrt{\tfrac{L_\star}{p_a\,\mu_a^2}}. \tag{27}$$

Proof. Fix $a \in [A]$ and $u \in \Omega_a$. Since $T(u) \in \operatorname{cone}_k(\mathbf{D})$ and the cone is invariant under positive scaling, $T(u)/\alpha_a(u) \in \operatorname{cone}_k(\mathbf{D})$, so

$$\|T(u) - \alpha_a(u)\mathbf{d}_a^\star\| = \alpha_a(u)\,\big\|T(u)/\alpha_a(u) - \mathbf{d}_a^\star\big\| \ge \alpha_a(u) \cdot \Delta_a, \tag{28}$$

where $\Delta_a := \operatorname{dist}(\mathbf{d}_a^\star, \operatorname{cone}_k(\mathbf{D}))$. By the triangle inequality and $\|r_a(u)\| \le \rho_a \alpha_a(u)$,

$$\|T(u) - u\mathbf{W}\| \ge \|T(u) - \alpha_a(u)\mathbf{d}_a^\star\| - \|r_a(u)\| \ge \alpha_a(u)(\Delta_a - \rho_a). \tag{29}$$

Using $\alpha_a(u) \ge \mu_a$ and taking positive parts,

$$\|T(u) - u\mathbf{W}\|^2 \ge \mu_a^2(\Delta_a - \rho_a)_+^2 \quad \text{for all } u \in \Omega_a. \tag{30}$$

Integrating over $\mathbb{S}^{n-1}$ and using $\sigma(\Omega_a) \ge p_a$,

$$L_\star \ge L_\sigma(T; \mathbf{W}) \ge p_a\,\mu_a^2\,(\Delta_a - \rho_a)_+^2, \tag{31}$$

which rearranges to (27). □

Remark L.10 (Informativity and finite probes). The bound is informative when $\rho_a + \sqrt{L_\star/(p_a\mu_a^2)}$ is well below the trivial value 1, since $0 \in \operatorname{cone}_k(\mathbf{D})$ and $\|\mathbf{d}_a^\star\| = 1$. The guarantee is most relevant for input regions with non-negligible solid angle p_a, sufficient write strength μ_a, and bounded off-axis residual ρ_a. Narrow regions or weak writes may have very small p_a or μ_a and therefore fall outside the practical regime of the bound. With m independent isotropic probes, the probability that Ω_a is never sampled is at most $(1 - p_a)^m$, and a union bound shows that sampling every one of A input regions with probability at least $1 - \delta$ is ensured by

$$m \gtrsim p_{\min}^{-1} \log(A/\delta), \qquad p_{\min} := \min_a p_a. \tag{32}$$

This sampling statement concerns coverage of the input regions rather than recovery of individual atoms, and it does not give an optimization guarantee for the nonlinear transcoder.

Remark L.11 (Scope of the coverage guarantee). Proposition L.9 bounds the distance from a reference direction to sparse nonnegative combinations of decoder rows. A small bound ensures that such a combination approximates the direction, but does not require an individual row to align with it or identify a unique dictionary. Token associations and code-level intervention responses are evaluated empirically in Appendix D.1 and Section 4.3 and are not implied by this geometric guarantee.

Remark L.12 (Role of isotropic probing). Assumption L.8 is a structural property of $\mathbf{W}$, not of the probe distribution. Isotropic Gaussian probes weight directional errors uniformly across $\mathbb{S}^{n-1}$ (Corollary L.2) and assign positive measure to every region of positive solid angle. Anisotropic probes that concentrate on a strict subset of the sphere can leave some Ω_a undersampled, weakly constraining the corresponding cone direction (Corollary L.4). Isotropic probing therefore assigns nonzero weight in the population loss to every input region of positive solid angle, independently of any semantic interpretation.