CROSS-ARCHITECTURE UNIVERSAL SAEs: A SHARED SPARSE DICTIONARY FOR VISION AND DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Universal Sparse Autoencoders (USAEs) learn a single sparse dictionary whose indices denote the same concept across several models, but the original USAE was evaluated only on vision transformers. We extend the USAE to a vision transformer (DINOv2) and a diffusion transformer (PixArt-α). We train a TopK dictionary with a separate encoder and decoder for each model, map both models' tokens to a common 16×16 patch grid, and add a per-token alignment loss that encourages the two models to activate the same indices at corresponding patches. On held-out COCO images, a code computed from either model explains variance in the other model's activations: $R^2 = 0.32$ from DINOv2 to PixArt-α and 0.14 in the reverse direction. At the average patch the two models share 43 of their 128 active indices, 38.7 times the rate expected under independent selection. Of the 12,288 dictionary indices, 74.7% are used by both models. Because the training objective rewards per-token agreement, the co-firing rate measures what the objective achieves, not whether the two architectures converge without it. We also show that agreement measures that pool codes over an image's patches are invariant to any permutation of one model's patches, so they cannot detect whether the models agree at corresponding locations. On our checkpoint, the pooled bag-of-features cosine is 0.9955, while the per-token Jaccard index is 0.203. We therefore report cross-model agreement per token and relative to a chance baseline.

1 INTRODUCTION

Mechanistic interpretability decomposes a network's activations into human-understandable features such as objects, parts, and textures (Olah et al., 2020). Networks appear to represent more features than they have neurons, so individual neurons respond to mixtures of unrelated concepts (Elhage et al., 2022). Sparse autoencoders (SAEs) address this superposition by learning an overcomplete dictionary in which each activation is reconstructed from a small number of active directions. In language models this recovers features that are more interpretable than individual neurons (Bricken et al., 2023; Cunningham et al., 2023; Templeton et al., 2024). SAEs have also been applied to vision transformers (ViTs) (Dosovitskiy et al., 2021) and CLIP-style models, where features can be inspected through image exemplars and manipulated at the patch level (Stevens et al., 2025; Zaigrajew et al., 2025).

These studies typically analyze one model at a time. A dictionary learned from one network does not show whether another network encodes the same concepts, so each new model needs its own analysis. If different models converge on similar features, as the circuits literature (Olah et al., 2020) and the Platonic Representation Hypothesis (Huh et al., 2024) suggest, then interpretability findings could transfer between models. Universal Sparse Autoencoders (USAEs) test this by training a single overcomplete SAE that encodes activations from any model in a pool and decodes them into the activations of every model, so that one dictionary index denotes the same concept in every model (Thasarathan et al., 2025).

The USAE was evaluated on DINOv2 (Oquab et al., 2023), SigLIP (Zhai et al., 2023), and a supervised ViT (Thasarathan et al., 2025). All three are vision transformers that tokenize images into patches and are trained to produce image representations, so features shared among them are a

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.

weaker test of universality than features shared across architectural families. Diffusion transformers (DiTs) (Peebles & Xie, 2023) are instead trained to denoise image latents conditioned on a timestep and, in text-to-image systems such as PixArt-α (Chen et al., 2023), on text. Intermediate diffusion activations carry semantic and correspondence information (Tang et al., 2023), and the hidden states of diffusion transformers are weakly aligned with DINOv2 features (Yu et al., 2025). To our knowledge, no shared sparse dictionary has been learned across the two families.

We introduce the Cross Sparse Autoencoder (X-SAE), an extension of the USAE to models of different architectures, and apply it to DINOv2 and PixArt-α on COCO images (Lin et al., 2014). DINOv2 produces a 16×16 grid of patch tokens and PixArt-α a 32×32 grid. We average-pool the latter to 16×16 so that each token position covers the same image region in both models. PixArt-α activations are extracted at a single denoising timestep, following Tang et al. (2023). We train a TopK SAE (Gao et al., 2025) with a separate encoder and decoder per model over a shared latent space. In addition to self- and cross-reconstruction, we add a per-token alignment loss that encourages both models to activate the same indices at corresponding patches.

Contributions.

1. **X-SAE.** We extend Universal Sparse Autoencoders to a vision transformer and a diffusion transformer, with a shared TopK dictionary, a common token grid, and a per-token alignment objective.
2. **Cross-architecture measurements.** On held-out images, cross-model reconstruction reaches $R^2 = 0.32$ (DINOv2 to PixArt-α) and 0.14 (reverse), and per-token co-firing is 38.7 times chance. Of the dictionary indices, 74.7% are used by both models, so the dictionary does not partition into per-model halves. Adding a dense pre-TopK alignment term raises co-firing by 37%, but 78.8% of indices then fall below the dead-feature threshold.
3. **A per-token evaluation of cross-model agreement.** We show that measures that pool codes over an image's tokens are invariant to permutations of one model's tokens, and that the bag-of-features cosine is near its ceiling (0.9955) on our checkpoint. We report a per-token Jaccard index with a chance baseline computed from the observed active counts.

2 Method

2.1 Shared Dictionary

We adopt the USAE formulation (Thasarathan et al., 2025), in which models share a dictionary index set but keep separate encoders and decoders, with a TopK activation (Gao et al., 2025). For each model, with encoder W_{enc}, decoder W_{dec}, and bias b_{pre}, applied per token:

$$z_{\text{pre}} = W_{\text{enc}}(x - b_{\text{pre}}), \qquad z = \text{TopK}(z_{\text{pre}}), \qquad \hat{x} = W_{\text{dec}}\, z + b_{\text{pre}}. \tag{1}$$

TopK keeps the K largest entries of z_{pre} and sets the rest to zero; no ReLU is applied. Index k addresses column k of every decoder, so a code produced by one model's encoder can be decoded into any model's activation space.

2.2 Token Correspondence

We average-pool PixArt-α's 32×32 token grid by a factor of two to DINOv2's 16×16 grid, so that token n covers the same image region in both models.

2.3 Training Objective

At each step a source model is chosen by step parity, alternating between the two models; USAE samples the source at random. The source code is decoded into both models' activation spaces, and the reconstruction loss is a weighted mean of the self- and cross-model squared errors, with weight 1 on the self term and 2 on the cross term.

Reconstruction does not constrain which indices each model activates: one model's decoder can learn to reconstruct its activations from indices that only the other model's encoder selects. In

USAE, cross-model alignment comes only from the shared reconstruction objective. We add an explicit term: one minus the cosine similarity between the two models' codes at each token position, averaged over positions. Both encoders are evaluated on the same image at every step, and both receive gradient from this term. We compare codes per token because a loss on token-pooled codes only matches the two models' image-level usage histograms, which the encoders can satisfy while activating disjoint indices at every position.

Under hard TopK, the alignment term passes no gradient to unselected indices. Within a step it adjusts shared indices toward agreement and shrinks indices selected by only one model; it cannot raise an index a model did not select. Shrinking an index can remove it from the top K at a later step, so over training the term does change which indices are selected.

A second alignment term, applied to the dense pre-activations z_{pre}, reaches every index. Its weight is zero in the main configuration; we ablate it in Section 4.5.

3 Experimental Setup

3.1 Models and Activation Extraction

DINOv2. We use DINOv2 ViT-S/14 (Oquab et al., 2023) on a 224-pixel center crop, taking the 256 final-layer normalized patch tokens (dimension 384) and discarding the CLS token. As in USAE, we use the final layer.

PixArt-α. We use PixArt-α XL/2 (1024-MS checkpoint) (Chen et al., 2023) on a 512-pixel center crop. The VAE encodes the crop to a 64×64 latent, which the transformer patchifies into 1,024 tokens of dimension 1,152. Following Tang et al. (2023), we noise the clean latent directly to a single timestep and run one transformer forward pass per noise draw. We use $t = 285$ of 1,000 training timesteps, index 10 of a 15-step DDIM schedule, which is close to the $t = 261$ that Tang et al. (2023) use for Stable Diffusion (Appendix A.1). Activations are averaged over eight noise draws, with a per-image seed so that extraction is deterministic. We take the output of transformer block 8 (zero-indexed) of 28, which includes the residual stream. REPA (Yu et al., 2025) obtains its best results by aligning the output of the eighth of 28 blocks of a diffusion transformer with DINOv2 features, and we hook at a comparable depth. We condition on the empty prompt with the T5 padding mask applied, so cross-attention attends to a single text token and adds the same vector at every token and image.

Both backbones receive the same aspect-preserving center crop at their respective resolutions before the pooling of Section 2.2. Activations are standardized per channel with statistics fitted on the training split after pooling. These statistics are stored with the checkpoint and reused at evaluation (Appendix A.3).

3.2 Data

We sample 2,000 images from COCO val2017 (Lin et al., 2014) with a fixed seed and split them 80/20 with a seeded shuffle into 1,600 training and 400 held-out images. The same selection and split are used for all caching, training, and evaluation.

3.3 Training Configuration

The dictionary has 12,288 indices, and TopK keeps $K = 128$ per token (1.04% density). The encoders and decoders have 37.8M parameters in total. The per-token alignment weight is 0.5. Cross-reconstruction and alignment are disabled for the first 5 of 30 epochs so that each encoder first learns per-model structure, before the alignment term can reward both encoders for selecting the same small set of indices on every input. We use the final (epoch 29, zero-indexed) checkpoint. Optimizer settings and dead-feature resampling are given in Appendix A.2.

3.4 Evaluation Metrics

Reconstruction. We report MSE for all four source–target pairs, and R^2 as $1 - \mathrm{MSE}$, since activations are standardized with training-split statistics.

Per-token co-firing. At each token position we compute the Jaccard index between the two models' sets of active indices and average over all positions and images. The chance baseline assumes that each model selects its active indices independently and uniformly at random, given the observed number of active indices at that position. The expected intersection is $n_A n_B/m$ for dictionary size m, and the baseline is the ratio of expected intersection to expected union, averaged over positions. With $n_A = n_B = 128$ and $m = 12{,}288$ this equals 0.005236. We report the lift, the ratio of the observed Jaccard index to this baseline.

Pooled measures. Each of the following reduces an image's token codes to one vector or set per model before comparing models. The *bag-of-features cosine* is the cosine between the two models' token-averaged $|z|$ vectors for an image. The *top-64 Jaccard* is the Jaccard index between each model's 64 features with the largest per-image maximum $|z|$. The *per-feature image co-occurrence* is, for each feature, the Jaccard index between the sets of images on which that feature is active (on any token) for each model. It is computed over features active in at least one model.

Dictionary occupancy. An index is *used* by a model if it is nonzero on at least one token of at least one image. The partition score is the larger model-exclusive count divided by the shared count. A value above 1 means the dictionary has split into per-model halves.

Top-image overlap. A feature's *peak activation* on an image is its maximum activation over tokens. For each model, we rank the held-out images by a feature's peak activation and keep the top five. The *image overlap* is the Jaccard index between the two models' top-five image sets. The *label overlap* is the Jaccard index between the sets of COCO categories annotated in each model's top-five images.

Reconstruction, co-firing, and top-image overlap are computed on the 400 held-out images. Pooled measures and occupancy are computed on all 2,000 images.

4 Results

4.1 Cross-Model Reconstruction

Table 1 reports all four source–target pairs. Self-reconstruction reaches R^2 of 0.773 for DINOv2 and 0.737 for PixArt-α. Both cross-model directions have positive R^2, so each model's code carries information about the other model's activations at the same patch. The directions are asymmetric: DINOv2 codes explain 32% of PixArt-α's variance, while PixArt-α codes explain 14% of DINOv2's. For reference, USAE reports off-diagonal R^2 between 0.25 and 0.45 among its three ViTs (Thasarathan et al., 2025). Our DINOv2-to-PixArt-α value falls within that range, and the reverse direction falls below it. The comparison is approximate, since the datasets, layers, dictionary sizes, and channel weighting of R^2 differ. The train–held-out gap is largest for DINOv2 to PixArt-α (0.125) and smallest for PixArt-α self-reconstruction (0.002).

Table 1: Reconstruction MSE for every source–target pair on held-out and training images. Activations are standardized with training-split statistics, so predicting the per-channel mean gives an MSE of approximately 1.

Source	**Target**	**Held-out**	**Train**	**Held-out R^2**
DINOv2	DINOv2	0.227	0.198	0.773
DINOv2	PixArt-α	0.682	0.557	0.318
PixArt-α	DINOv2	0.861	0.770	0.139
PixArt-α	PixArt-α	0.263	0.261	0.737

4.2 Per-Token Co-Firing

The observed per-token Jaccard index is 0.2028, 38.7 times the chance value of 0.005236 (Table 2, main configuration). With 128 active indices per model, this corresponds to 43 shared indices at the average token. Because the objective penalizes per-token disagreement (Section 2.3), this result shows that the objective achieves its target. It does not show how much agreement would arise from reconstruction alone.

Table 2: Per-token co-firing and held-out reconstruction MSE with and without the dense pre-TopK alignment term (Section 4.5). The per-token alignment term on the TopK codes is active in both configurations. "Below dead-feature threshold" counts indices whose firing-rate EMA is below 10^{-3} in both models at the final checkpoint (Appendix A.2); it is not comparable to Table 3.

	Main	**+ Pre-TopK align**
Per-token Jaccard	0.2028	0.2777
Chance	0.0052	0.0052
Lift over chance	38.7	53.0
DINOv2 $\rightarrow$ PixArt-α	0.682	0.665
PixArt-α $\rightarrow$ DINOv2	0.861	0.851
DINOv2 $\rightarrow$ DINOv2	0.227	0.273
PixArt-α $\rightarrow$ PixArt-α	0.263	0.283
Below dead-feature threshold	1,120 (9.1%)	9,689 (78.8%)

4.3 Dictionary Occupancy

Three quarters of the dictionary is used by both models, and the partition score of 0.224 is well below 1 (Table 3). DINOv2 has 4.5 times as many exclusive indices as PixArt-α, and it is also the model that is harder to predict from the other's code (Table 1). USAE likewise finds that DINOv2 encodes model-specific concepts and has the lowest co-fire proportion of the three ViTs it studies (Thasarathan et al., 2025). Occupancy is a pooled measure (Section 4.4): it shows that the dictionary has not partitioned, but not that shared indices fire at the same locations.

Table 3: Usage of the 12,288 dictionary indices over all 2,000 images. The partition score is $2{,}055/9{,}174 = 0.224$.

Used by	**Count**	**Share**
Both models	9,174	74.7%
DINOv2 only	2,055	16.7%
PixArt-α only	458	3.7%
Neither	601	4.9%

4.4 Pooled Measures Cannot Detect Spatial Correspondence

The pooled measures in Table 4, like occupancy and USAE's firing entropy, reduce each model's codes over tokens before comparing models: they average, take a maximum, or record whether a feature fires anywhere in the image. None of these reductions depends on token order. Because each encoder acts on tokens independently, permuting one model's tokens within each image permutes its codes in the same way. Such a permutation destroys any correspondence between patches, yet leaves every pooled measure unchanged, while the per-token Jaccard index changes. Pooled measures therefore cannot distinguish a dictionary whose shared indices fire at the same patches from one whose shared indices fire at unrelated patches. USAE's co-fire proportion is computed per patch (Thasarathan et al., 2025) and does not have this limitation. Our per-token Jaccard index is a set-level analogue of it with a chance baseline.

On our checkpoint the bag-of-features cosine is 0.9955, close to its maximum of 1. The per-token Jaccard index is 0.203 and changes by 37% between the two configurations of Table 2. The per-feature image co-occurrence has a mean of 0.101 and a first quartile of 0, so at least a quarter of the 11,687 active features (at least 2,922) never fire for both models on the same image. Of these, 2,513 are model-exclusive and have co-occurrence 0 by definition, so at least 409 features used by both models never fire for both on the same image. The per-feature measure weights every active feature equally, including model-exclusive and rarely active ones, whereas the per-token measure weights each feature by how often it is active.

Table 4: Agreement measures for the main configuration. The first three pool over tokens and are invariant to any permutation of one model's tokens. Pooled measures use all 2,000 images; the per-token Jaccard index uses the 400 held-out images.

Measure	Permutation-invariant	Mean	Median
Bag-of-features cosine (per image)	yes	0.9955	0.9961
Top-64 Jaccard (per image)	yes	0.667	0.662
Per-feature image co-occurrence	yes	0.101	0.051
Per-token Jaccard	no	0.203	—

4.5 Effect of Pre-TopK Alignment

Adding the dense term raises the per-token Jaccard index by 37% and lowers cross-model reconstruction error by at most 0.017 (Table 2). Self-reconstruction error rises by 0.046 for DINOv2 and 0.020 for PixArt-α, and the share of indices below the dead-feature threshold rises from 9.1% to 78.8%. With 128 indices active at every token, the two models thus reach higher per-token agreement while concentrating their activity on roughly a fifth of the dictionary. Because the term costs self-reconstruction and most of the dictionary for a small change in cross-model reconstruction, we set its weight to zero in the main configuration. The per-token term on the TopK codes is active in both runs, so this ablation does not isolate the contribution of that term.

4.6 Qualitative Examples

Figure 1 shows the top-five held-out images of feature 1091 for each model. Zebra is the only COCO category annotated in any of the ten images, four images appear in both models' top five, and in both models the activation is concentrated on the zebras. Appendix B reports the top-image overlap for the twenty features with the highest peak activation on held-out images. Across these features the mean image overlap is 0.337 and the mean label overlap is 0.895, and label overlap exceeds image overlap for every feature (Table 5). The two models' top images for a shared index usually differ but are annotated with the same categories. Distinct COCO images often share categories such as person, so label overlap can be high even for unrelated image sets, and we do not compare it against a baseline of random image sets.



Figure 1: The five held-out images with the highest peak activation of feature 1091 for DINOv2 (top) and PixArt-α (bottom). Overlays show the feature's per-token activation on the 16×16 grid. Labels give the last six digits of the COCO image ID and the peak activation. Four of the five images appear in both rows (image overlap 0.67), and the sets of COCO categories annotated in the two rows are identical (label overlap 1.00).

5 Related Work

Sparse autoencoders for interpretability. Decomposing representations into sparse parts builds on sparse coding and dictionary learning (Olshausen & Field, 1996). The superposition hypothesis attributes neuron polysemanticity to networks encoding more features than they have dimensions (Elhage et al., 2022). Yun et al. (2021) applied sparse dictionary learning to transformer representations, and overcomplete SAEs were then shown to recover language-model features that are more monosemantic than those found by alternative methods (Cunningham et al., 2023; Bricken et al., 2023). SAEs scale to production language models (Templeton et al., 2024). Later work replaced the L_1 penalty with JumpReLU (Rajamanoharan et al., 2024) and TopK activations (Gao et al., 2025). We use TopK because it fixes the number of active features per token, which makes active sets directly comparable across models. Other work addresses feature absorption with hierarchical dictionaries (Bussmann et al., 2025), finds that SAE proxy metrics do not reliably predict downstream performance (Karvonen et al., 2025), and questions whether SAE features are canonical units of analysis (Leask et al., 2025). Most of this literature studies one model at a time.

Universal and cross-model representations. Model stitching showed that an intermediate representation of one network can be mapped into another through a learned affine layer (Lenc & Vedaldi, 2015). Later work showed that self-supervised and supervised models can be stitched with little loss in performance (Bansal et al., 2021) and that stitching is possible without learning a mapping (Moschella et al., 2023). Rosetta Neurons match units across discriminative and generative vision models, including GANs (Dravid et al., 2023), and related work finds shared concepts in video transformers (Kowal et al., 2024). The Platonic Representation Hypothesis argues that representations converge across architectures and modalities as models scale (Huh et al., 2024). These analyses compare independently learned representations after training, whereas sparse dictionary methods learn a shared representation directly. Crosscoders learn dictionaries spanning layers and support model diffing (Lindsey et al., 2024), and USAEs learn one dictionary that encodes and decodes activations of several models (Thasarathan et al., 2025). We build on USAEs, in which alignment across models comes only from the shared reconstruction objective. SPARC (Nasiri-Sarvi et al., 2026) adds explicit concept alignment across models and modalities; we add an explicit per-token alignment term for spatially corresponding patches. USAE and SPARC were evaluated on encoder models only; neither includes a generative model.

Interpretability of vision and diffusion models. Self-supervised ViTs such as DINOv2 produce spatially precise patch representations (Oquab et al., 2023), and SAEs trained on vision backbones recover concepts that support patch-level interventions (Stevens et al., 2025). USAE finds that DINOv2 encodes concepts not shared with SigLIP or a supervised ViT (Thasarathan et al., 2025). Work on diffusion models first targeted U-Nets, identifying semantic directions in the bottleneck layer (Park et al., 2023), and SAEs have been trained on the transformer blocks of a text-to-image U-Net, where their features causally affect generation (Surkov et al., 2025). Diffusion transformers (Peebles & Xie, 2023; Chen et al., 2023) replace the U-Net's multi-resolution hierarchy with a single-resolution token sequence. TIDE trains timestep-aware SAEs on DiTs and shows that their representations vary across the denoising schedule (Huang et al., 2025). REPA shows that diffusion transformer hidden states are weakly aligned with DINOv2 features and that strengthening this alignment during training improves generation (Yu et al., 2025). REPA operates on dense representations and does not learn a sparse dictionary. Existing SAE work on diffusion models interprets each model with its own dictionary; we interpret a representation model and a generative model through one shared dictionary.

6 Limitations and Conclusion

We trained a TopK sparse autoencoder with one dictionary jointly on DINOv2 and PixArt-α. On held-out COCO images, cross-model reconstruction reaches $R^2 = 0.32$ from DINOv2 to PixArt-α and 0.14 in the reverse direction, and the two models share 43 of 128 active indices at the average patch, 38.7 times chance. Of the 12,288 dictionary indices, 74.7% are used by both models. We also show that pooled agreement measures are invariant to permutations of one model's tokens, so cross-model agreement should be measured per token.

Because the objective penalizes per-token disagreement, the co-firing results do not show that the two architectures converge without that objective; separating the two requires a run with the alignment weight set to zero. We do not compare against SAEs trained independently on each model and matched after training. Shared indices firing at the same patch also do not show that they encode the same visual concept; our evidence on this is limited to the top-image overlap of twenty features, without a random baseline or causal tests. The study uses one model pair, one transformer block, one denoising timestep, one sparsity level, one dictionary size, and 2,000 images. PixArt-α is run at 512 pixels, below the 1024-pixel resolution of the checkpoint we use.

Ethics Statement

All experiments use COCO (Lin et al., 2014), a public dataset. We collected no new data and ran no human-subject studies. COCO contains identifiable people, and dictionary features learned on it can localize person-related content. The method does not modify either backbone, but a shared dictionary that indexes a generative model's activations could be used to steer generation as well as to audit it.

LLM Usage Statement

LLM usage was limited to the following: help with literature review; assistance in drafting the Related Work section, which was subsequently edited and verified by the authors; grammar checks throughout the paper; and generating code snippets that were verified by the authors.

References

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, 2021.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. `https://transformer-circuits.pub/2023/monosemantic-features/index.html`.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with Matryoshka sparse autoencoders. *arXiv preprint arXiv:2503.17547*, 2025.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, et al. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Amil Dravid, Yossi Gandelsman, Alexei A. Efros, and Assaf Shocher. Rosetta neurons: Mining the common units in a model zoo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, et al. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, 2025.

Victor Shea-Jay Huang, Le Zhuo, Yi Xin, Zhaokai Wang, Fu-Yun Wang, Yuchi Wang, Renrui Zhang, Peng Gao, and Hongsheng Li. TIDE: Temporal-aware sparse autoencoders for interpretable diffusion transformers in image generation. *arXiv preprint arXiv:2503.07050*, 2025.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, et al. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *International Conference on Machine Learning*, 2025.

Matthew Kowal, Achal Dave, Rares Ambrus, Adrien Gaidon, Konstantinos G. Derpanis, and Pavel Tokmakov. Understanding video transformers via universal concept discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *International Conference on Learning Representations*, 2025.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, et al. Microsoft COCO: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing. *Transformer Circuits Thread*, 2024. https://transformer-circuits.pub/2024/crosscoders/index.html.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations*, 2023.

Ali Nasiri-Sarvi, Hassan Rivaz, and Mahdi S. Hosseini. SPARC: Concept-aligned sparse autoencoders for cross-model and cross-modal interpretability. *Transactions on Machine Learning Research*, 2026.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. https://distill.pub/2020/circuits/zoom-in/.

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Yong-Hyun Park, Mingi Kwon, Junghyo Jo, and Youngjung Uh. Unsupervised discovery of semantic latent directions in diffusion models. *arXiv preprint arXiv:2302.12469*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024.

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

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 *Advances in Neural Information Processing Systems*, 2025.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Advances in Neural Information Processing Systems*, 2023.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos G. Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment. In *International Conference on Machine Learning*, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

Zeyu Yun, Yubei Chen, Bruno A. Olshausen, and Yann LeCun. Transformer visualization via dictionary learning: Contextualized embedding as a linear superposition of transformer factors. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, 2021.

Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. Interpreting CLIP with hierarchical sparse autoencoders. In *International Conference on Machine Learning*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

A Implementation Details

A.1 Timestep Selection

We selected the PixArt-α timestep with a sweep over all 15 timesteps of the DDIM schedule, run on an earlier checkpoint trained before the timestep was fixed at index 10. For each timestep we counted active dictionary features and measured peakiness, the ratio of the maximum to the mean per-token $|z|$. At index 14 ($t = 0$), 288 features are active and peakiness is 1.22. At index 10 ($t = 285$), 3,617 features are active and peakiness is 6.09. Peakiness does not depend on token order and carries no spatial information, so we used the sweep only to exclude timesteps at which the diffusion representation is nearly uniform across tokens.

A.2 Optimization and Dead-Feature Resampling

Each decoder is initialized as the transpose of its encoder, and decoder columns are renormalized to unit norm after every step. We train for 30 epochs with AdamW, batch size 16, learning rate 5×10^{-4} with 50 warmup steps and cosine decay to 1×10^{-6}, and weight decay 1×10^{-5}.

Every 501 steps, up to 512 dead indices are resampled. The interval is odd so that resampling events alternate between source models. An index is dead when its firing-rate EMA is below 10^{-3} in both models. For each model we sample a token with probability proportional to its squared self-reconstruction error, set the index's decoder column to the normalized residual at that token, set its encoder row to the same direction scaled to 0.05 times the mean norm of live encoder rows, and reset its optimizer state.

A.3 Normalization

Per-channel means and standard deviations are computed after pooling over 1,000 images from the training split and stored with the checkpoint. At the encoder input, the measured standard deviation is 0.998 for DINOv2 and 1.000 for PixArt-α.

B TOP-IMAGE OVERLAP OF SHARED FEATURES

We select the twenty features with the highest peak activation on held-out images; the selection does not use either overlap measure. For each feature we compute the image overlap and label overlap defined in Section 3.4. Table 5 lists the twenty features sorted by label overlap, and Figure 2 shows four of them. Label overlap exceeds image overlap for all twenty features: the two models' top images for a feature usually differ but are annotated with the same categories. Four of the twenty features (1091, 1760, 10487, and 11092) have zebra as their only shared category. Multiple features responding to one concept is consistent with feature splitting (Bricken et al., 2023); we do not test whether these four features respond to different aspects of zebra images.

Table 5: Top-image overlap between DINOv2 and PixArt-α for the twenty features with the highest peak activation on held-out images, sorted by label overlap. "Shared labels" lists up to four of the COCO categories annotated in both models' top-five images, in alphabetical order.

#	Feature	Image overlap	Label overlap	Shared labels
1	1091	0.667	1.000	zebra
2	1760	0.250	1.000	zebra
3	3545	0.111	1.000	person, skis, snowboard
4	8767	0.667	1.000	chair, person, sports ball, tennis racket
5	8833	0.429	1.000	airplane, boat, bus, car
6	9898	0.250	1.000	backpack, car, kite, person
7	10487	0.429	1.000	zebra
8	10499	0.000	1.000	person, surfboard
9	11092	0.429	1.000	zebra
10	9359	0.250	0.857	backpack, baseball bat, baseball glove, person
11	9790	0.250	0.857	backpack, baseball bat, baseball glove, cup
12	11457	0.429	0.857	backpack, baseball glove, chair, person
13	92	0.429	0.818	bowl, broccoli, carrot, cup
14	1073	0.250	0.800	airplane, person, train, truck
15	3348	0.250	0.800	car, kite, person, snowboard
16	3720	0.111	0.800	chair, person, sports ball, tennis racket
17	2347	0.429	0.778	airplane, bicycle, bottle, chair
18	4250	0.429	0.778	bear, bottle, cow, horse
19	4792	0.250	0.778	bench, bicycle, bus, car
20	5921	0.429	0.778	backpack, baseball bat, baseball glove, bench
Mean		0.337	0.895	

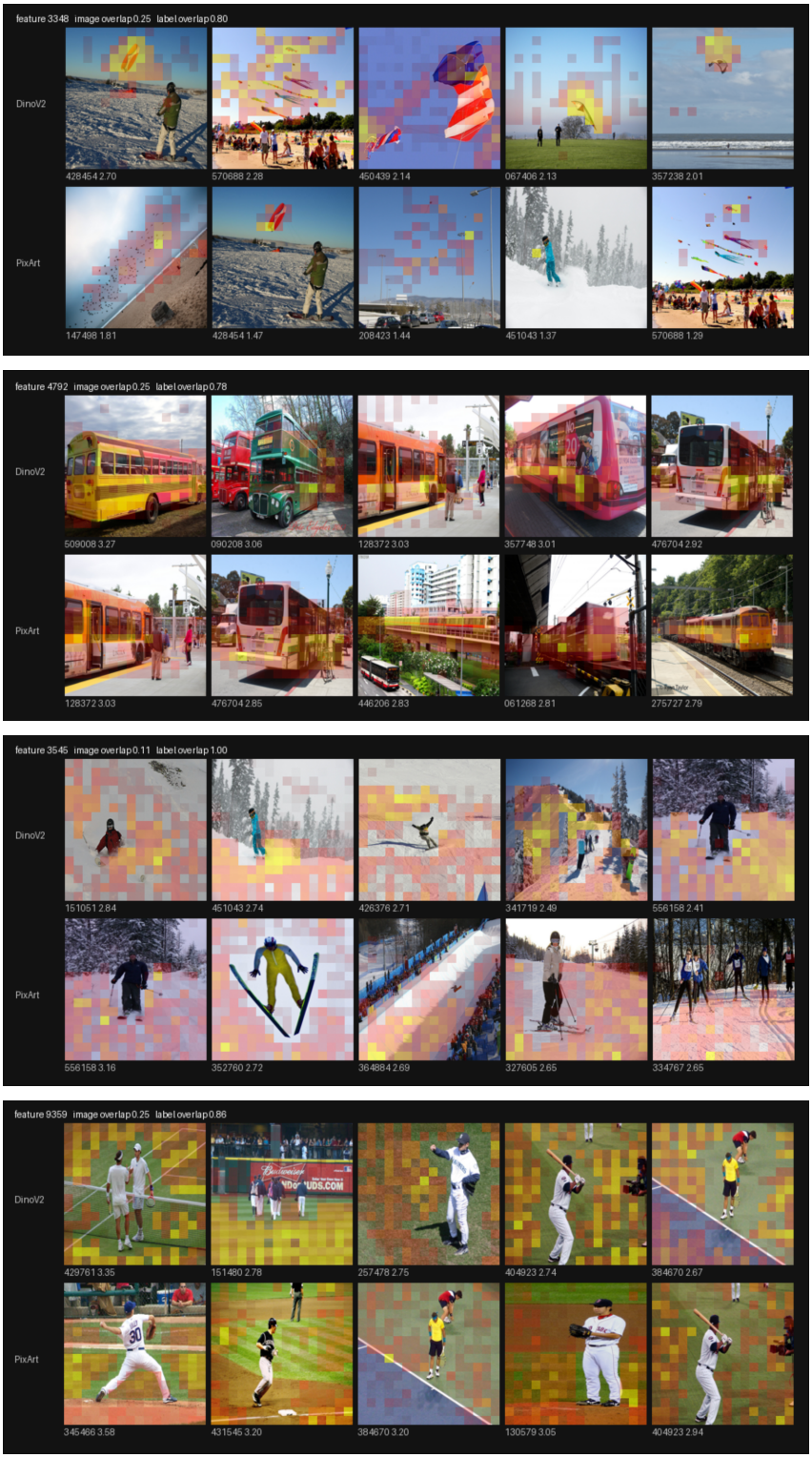

Figure 2: Top-five held-out images for DINOv2 (top row of each panel) and PixArt-α (bottom row) for features 3348, 4792, 3545, and 9359, with per-token activation overlays. The features activate on scene backgrounds such as snow, courts, and fields, and on objects such as kites and buses.