AUTOMATIC INTERPRETATION OF VISUAL CONCEPTS

Felix Meissen*, **Kenza Bouzid, Tristan Lazard, Shruthi Bannur,**
Daniel C. Castro, Javier Alvarez-Valle, Stephanie Hyland
Microsoft Research
Cambridge, UK

ABSTRACT

Recent progress in mechanistic interpretability and sparse autoencoders (SAEs) has opened new avenues for understanding vision models, yet automatically assigning accurate textual descriptions to discovered concepts remains unprincipled. Existing studies rely on proxy metrics such as CLIP similarity or qualitative inspection, which fail to measure semantic faithfulness of the concept descriptions. To bridge this gap, we conduct a principled study of the automatic interpretation pipeline, evaluating key design choices including MLLM query construction and sample selection. We introduce Semantic Label Quality (SLQ) metrics from language model interpretability to vision, providing direct measurement of label faithfulness. We further investigate whether synthetic counterfactuals generated by a conditional generative model can further improve interpretation. Experiments on synthetic faces, histopathology, and remote sensing images reveal that optimal interpretation strategies are dataset-dependent: no single configuration universally outperforms others. Counterfactual contrastive samples improve interpretation for localized, additive concepts but provide limited benefit for global concepts where counterfactuals are less well defined.

1 INTRODUCTION

Vision foundation models contain hundreds of internal features that align with human-understandable factors of variation (Pach et al., 2025), and sparse autoencoders (SAEs) can isolate these features at scale. Yet assigning accurate textual descriptions to features—and verifying them—remains ad hoc. Current practice treats automated labelling mainly as a proxy for the interpretability of detected features, evaluating labels qualitatively, or with CLIP similarity or BERTScore (Hernandez et al., 2022; Oikarinen & Weng, 2023)—none of which guarantee that a label faithfully captures the *semantics* of the underlying concept at scale.

Automated interpretation is challenging because high-activation image sets contain co-varying attributes, and negatives are poorly controlled. Prompting a multimodal large language model (MLLM) to explain a concept from highly activating samples invites spurious correlations. Design decisions, such as sample selection, prompt construction, or sample count, strongly affect outcomes but have not been systematically studied, and the field lacks reliable ways to test whether descriptions are *factually correct* for open-ended concepts, rather than merely fluent or plausible (Oikarinen & Weng, 2023; Zaigrajew et al., 2025).

One promising direction to control for co-varying attributes is synthesising contrastive pairs via conditional generative models. By decoding steered embeddings that differ only in the target concept, the resulting counterfactual pairs should isolate that concept from confounders. Recent work has explored this idea for explaining supervised classifiers (Zablocki et al., 2024) and text-to-image diffusion models (Cammarata et al., 2025; Olson et al., 2025; Surkov et al., 2025), yet rigorous evaluation of this approach for automatic interpretation of vision models is lacking.

In this paper, we address these gaps with a principled study of the automatic interpretation pipeline for vision concepts. Given a concept representation (*e.g.*, an SAE feature) and a dataset, the task of automatic interpretation is to produce a concise textual description that reliably predicts the presence of the concept. We evaluate the impact of: (1) MLLM query format (how many samples,

*Corresponding author: felixmeissen@microsoft.com

how to arrange them), and (2) sample selection strategies, including counterfactual generation via a conditional generative model. We build confidence in our results by introducing semantic label quality (SLQ) metrics, previously used for language models (Paulo et al., 2025), to the vision domain. These metrics provide standard measures (*e.g.* precision, recall) of the consistency of proposed labels with the underlying features. Our contributions are as follows:

- **Principled automatic interpretation study:** We isolate design choices in the automatic interpretation pipeline and find that optimal settings are dataset-dependent. We validate findings across proprietary (GPT-4.1) and open-weights (Qwen3-VL) MLLMs.
- **Semantic Label Quality (SLQ) metrics for vision:** We introduce concept detection scores to vision, defining SLQ as consistency between MLLM classifications and concept activations.
- **Investigation of counterfactual interpretation:** We investigate to what extent counterfactual pairs generated by a conditional generative model (CONCEPTDECODER) can improve concept descriptions by reducing confounders.

2 RELATED WORK

Mechanistic interpretability with SAEs. Since their introduction by Bricken et al. (2023), sparse autoencoders (SAEs) have been applied at scale to language models (Templeton et al., 2024; Gao et al., 2024; Lieberum et al., 2024; He et al., 2024) and extended to vision (Stevens et al., 2025; Pach et al., 2025; Olson et al., 2025), multimodal models (Zhang et al., 2025b), and medical imaging (Le et al., 2024; Abdulaal et al., 2024; Bouzid et al., 2025; Nakka, 2025).

Interpretation of image concepts. Early work interpreted CNN neurons by matching them to prelabelled concepts in manually curated datasets (Bau et al., 2017). Subsequent approaches use CLIP to match highly activating samples to descriptions from predefined concept banks (Oikarinen & Weng, 2023; Zaigrajew et al., 2025; Rao et al., 2024; Kalibhat et al., 2023). Moving beyond fixed label sets, MILAN (Hernandez et al., 2022) trains a neuron captioning model on a custom dataset of fine-grained image annotations. Recent work trains SAEs on text-to-image diffusion models and steers generation for label assignment (Cammarata et al., 2025; Olson et al., 2025; Surkov et al., 2025), though this is limited to evaluating the generative model itself. Hoang-Xuan et al. (2024) proposed a multi-step process of selecting highly activating samples, filtering, and asking an MLLM to label the cluster. Zablocki et al. (2024) generate counterfactual pairs for explaining supervised classifiers but do not evaluate explanation correctness against available ground truth, nor do they provide experiments on how to obtain high-quality explanations. Crucially, rigorous evaluation remains absent: current approaches rely on qualitative inspection (Zaigrajew et al., 2025; Zablocki et al., 2024), or proxy metrics like BERTScore and CLIP similarity (Hernandez et al., 2022; Oikarinen & Weng, 2023; Olson et al., 2025; Surkov et al., 2025). While some works treat interpretation as classification with known concepts (Oikarinen & Weng, 2023; Rao et al., 2024; Hoang-Xuan et al., 2024), there is currently no reliable automated method to verify if a proposed open-ended concept label is faithful to the underlying model.

Visualizing concepts with conditional generative models. Carballo-Castro et al. (2024) developed a counterfactual generation approach based on concept activations vectors, albeit limited to a fixed, known set of concepts. Cohen et al. (2021) created class-counterfactuals for chest X-rays as an alternative to saliency maps for explaining neural network decisions. Žigutyte et al. (2024) proposed a method to determine which morphological features in histopathology images drive classifier predictions by creating counterfactual explanations, while Nguyen et al. (2025) simulated morphological changes between classes to generate counterfactuals to help label features of interest.

3 METHODS

3.1 PROBLEM FORMULATION

Let $f_\theta : \mathcal{X} \to \mathcal{Z}$ be a vision foundation model encoding images $\mathbf{x} \in \mathcal{X}$ into representations $\mathbf{z} = f_\theta(\mathbf{x}) \in \mathcal{Z}$. We assume access to a *concept activation function* $a_c : \mathcal{Z} \to \mathcal{A}$ that quantifies how strongly concept c is expressed in a representation, where $\mathcal{A} \subseteq \mathbb{R}$. Such functions arise naturally from interpretability tools such as SAEs and linear probes, as we describe in Appendix B. Given a_c, a dataset $\mathbf{X}$, and an MLLM $\mathcal{M}$, the *automatic interpretation task* is to produce a textual description or

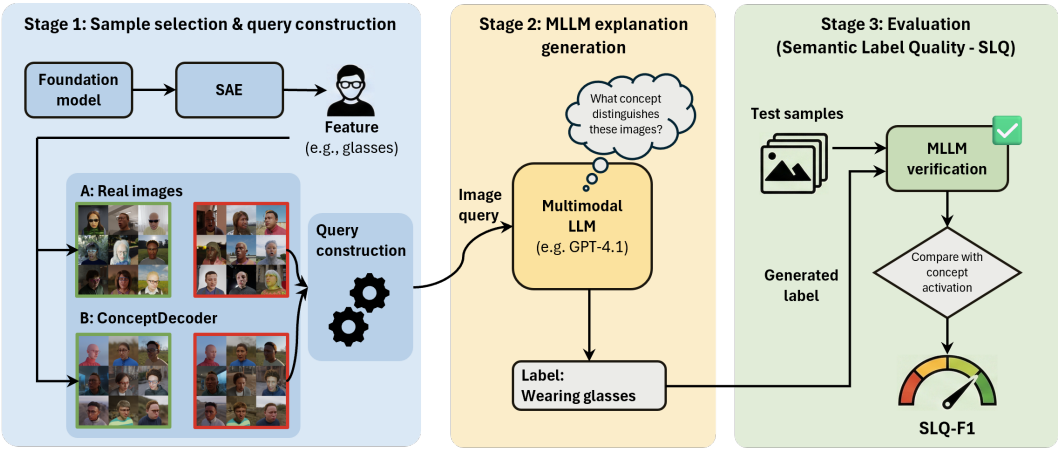


Figure 1: The end-to-end automatic interpretation pipeline. Stage 1 selects real or generates synthetic samples to isolate a specific feature and constructs a query. Stage 2 uses an MLLM to generate a textual description of the concept. Stage 3 evaluates the description's faithfulness using SLQ metrics by prompting the MLLM to classify held-out images based on the generated label.

label ℓ_c that faithfully captures the semantics of c. A description is *faithful* if it reliably corresponds to concept activation: images where the description applies should have high a_c, and vice versa.

While our setup applies to any concept activation function, we focus on *linear* concepts: SAE features and linear probes. For SAEs, we use encoder activations a_c^{enc} for sample selection (with binarization $a_c^+(\mathbf{z}) = \mathbb{I}[a_c^{\text{enc}}(\mathbf{z}) > 0]$) and decoder projections a_c^{dec} for steering (Bricken et al., 2023). For supervised concepts, $a_c(\mathbf{z}) = \langle \mathbf{v}_c, \mathbf{z} \rangle$ with threshold τ_c chosen for optimal F_1. See Appendix B for details.

3.2 Automatic interpretation pipeline

Our automatic interpretation pipeline consists of two stages: (1) *sample selection* (which images to show the MLLMs); and (2) *query construction* (how to arrange and present them). Each stage admits design choices that have not been systematically compared.

Sample selection. Modern MLLMs process only limited images per query effectively (OpenAI, 2025), making sample selection critical. We study two aspects:

Activation-based selection determines which activation range to draw samples from within the positive or negative samples. We base this on quantiles of the activation distribution $A_c^{\text{ref}} = \{a_c(\mathbf{z}) : \mathbf{z} \in \mathbf{Z}^{\text{ref}}\}$ over a reference set $\mathbf{Z}^{\text{ref}}$:

$$a_c^{\text{lo}} = \text{quantile}(\{a \in A_c^{\text{ref}} : a \geq \tau_c\}; q^{\text{lo}}), \quad a_c^{\text{hi}} = \text{quantile}(\{a \in A_c^{\text{ref}} : a \geq \tau_c\}; q^{\text{hi}}), \quad (1)$$

where $q^{\text{lo}} \leq q^{\text{hi}}$ define the quantile range. Samples are then selected from $\mathbf{X}$ with activations in $[a_c^{\text{lo}}, a_c^{\text{hi}}]$.

Negative sample selection determines how to obtain contrastive samples: (i) *Random*: negatives drawn from low-activation samples; (ii) *Nearest-neighbour*: retrieve real images closest to the steered representations (see Appendix C); (iii) *Counterfactual via* ConceptDecoder: generate synthetic pairs differing only in target concept (see below).

Query construction. Given selected positive and negative samples, we must decide how to present them to the MLLM. We study two decisions: the *number of exemplars* to include, and the *image arrangement* strategy. We experiment with five image presentation strategies for automatic interpretation, adhering to the technical restrictions of the model (*cf*. Appendix H.1) (OpenAI, 2025):

- `TwoGrids`: Positives and negatives combined into separate grid images (max 8×3 for 256×256 images).
- `OneGrid`: Positives in top row, negatives in bottom row, combined into one grid.

- `Paired`: Each positive-negative pair concatenated horizontally.
- `Separate`: Positive and negative images interleaved as separate images.

Counterfactual generation via CONCEPTDECODER. To generate a counterfactual pair, we steer each seed representation toward high and low activation targets along the concept direction, then decode using a CONCEPTDECODER (a conditional diffusion model; Section 4.4). See Appendix B for details on steering.

3.3 SEMANTIC LABEL QUALITY (SLQ) METRICS

We adapt the *detection score* from language model interpretability (Paulo et al., 2025) to evaluate the quality of concept descriptions and our design choices. Given a proposed description ℓ_c for concept c, we evaluate its faithfulness by testing whether an MLLM can use ℓ_c to predict binary concept activations on held-out images. For a set of query images $\{\mathbf{x}_j\}$, we prompt the MLLM to determine the presence of the concept ℓ_c in each image (see Appendix H.2 for the full prompt). Comparing the MLLM's binary predictions $\hat{y}_j$ against the ground-truth binarized activations $y_j = a_c^+(f_\theta(\mathbf{x}_j))$ yields SLQ-acc $= \frac{1}{N}\sum_j \mathbb{I}[\hat{y}_j = y_j]$ and SLQ-F_1 $= F_1(\{\hat{y}_j\}, \{y_j\})$. Unlike proxy metrics such as CLIP similarity, SLQ directly measures the consistency of a concept label.

4 EXPERIMENTAL SETUP

4.1 DATASETS

We conduct experiments on three complementary domains: **SynthFace** (Wood et al., 2021), a synthetic face dataset with rich metadata to extract labels for fine-grained facial attributes, enabling validation of our pipeline and of the SLQ metrics themselves; **Hancock** (Dörrich et al., 2025), histopathology images for precision oncology; and **BigEarthNet** (Clasen et al., 2025), a large-scale remote sensing dataset. This selection evaluates a broad spectrum of real-world concepts, from subtle and localized changes in human imagery to large-scale complex patterns in histopathology and satellite images. Dataset splits and processing details are in Appendix A.

4.2 CONCEPT EXTRACTION

We consider supervised and unsupervised settings. In the supervised setting, we train linear probes to predict SynthFace attributes; the weight vectors serve as concept directions. In the unsupervised setting, we derive concepts via Matryoshka BatchTopK SAEs (Bussmann et al., 2025; 2024). Details are in Appendices D and E.

4.3 AUTOMATED INTERPRETATION

We use GPT-4.1 (Azure OpenAI) as the primary MLLM, with Qwen3-VL-8B-Instruct (Yang et al., 2025) confirming generalization to open-weights models. SLQ evaluation uses balanced sets of up to 100 positive and 100 negative samples per concept, with GPT-4.1 as the judge due to its strong visual capabilities. Balancing reduces token costs and ensures SLQ-F_1 of 0.5 corresponds to random performance, facilitating comparisons. The subsampling introduces stochasticity; we measured a standard deviation of 0.33 percentage points across 5 runs, yielding a minimum detectable difference (MDD) of 0.6 pp at $p < 0.05$ (see Appendix G.1). For Hancock and BigEarthNet, the MDDs are 0.9 pp and 0.5 pp, respectively.

4.4 MODELS

Foundation models and concept space. For SynthFace and BigEarthNet, we use DINOv3 (Siméoni et al., 2025) as the vision backbone; generalization to other models is shown in Appendix J. For Hancock, we use UNI2 (Chen et al., 2024), a vision foundation model specifically designed for histopathology images. We use only the `CLS` token (or its architecture-specific global token) as the concept space, since this representation is commonly consumed in downstream tasks.

CONCEPTDECODER architecture and training. We implement CONCEPTDECODER as a conditional diffusion model. For SynthFace and BigEarthNet, we use a convolutional U-Net with a

"simple diffusion"-style training objective and pixel-shuffle operations (Hoogeboom et al., 2023) and cosine learning-rate schedule, generating 256×256 images. Training takes 378,000 steps ($\sim$3.5 days on 4×A100 GPUs); metrics are in Appendix F. For Hancock, we use the publicly available PixCell-256 (Yellapragada et al., 2025) trained to invert UNI2 embeddings (Chen et al., 2024).

5 RESULTS

Table 1: SLQ-F_1 for query construction and activation threshold experiments in %. BEN: BigEarthNet. "Extreme" selects the most extremely activated samples. "Top-k Sim" is the method from Hoang-Xuan et al. (2024). °Default setting used in (b).

(a) Query Construction

	SynthFace		Hancock	BEN
	GPT	Qwen	GPT	GPT
Num. samples (M):				
1	48.2	48.2	64.3	63.4
4	57.0	53.4	66.8	68.2
8	59.3	57.5	67.9	68.7
16	59.7	59.5	67.9	69.0
24°	**60.9**	**59.9**	**68.1**	**69.7**
Prompting strategy:				
`TwoGrids`°	**59.9**	**59.0**	67.9	69.4
`OneGrid`	57.3	54.3	66.2	69.4
`Paired`	56.9	50.1	65.8	70.4
`Separate`	56.1	50.5	**69.6**	**71.1**

(b) Activation Threshold

	SynthFace		Hancock	BEN
q^{lo}	GPT	Qwen	GPT	GPT
10	59.3	58.6	68.8	70.8
20	61.0	61.1	68.8	70.6
30	60.7	61.2	68.9	70.9
40	61.2	62.4	**69.9**	70.8
50	62.0	64.0	69.5	**71.2**
60	63.2	65.0	68.5	70.3
70	63.6	66.1	68.2	70.2
80	**65.0**	66.0	68.3	69.7
90	64.9	**66.4**	68.3	67.9
Extreme	64.1	64.6	63.6	65.0
Top-k Sim	64.2	64.1	63.5	64.5

Table 2: Experiments for negative sample selection strategies. All experiments were performed with the best settings from Table 1. We report the average SLQ-prec (SLQ-p), SLQ-rec (SLQ-r), and SLQ-F_1 across all concepts, in %.

	SynthFace						Hancock			BigEarthNet		
	GPT-4.1			Qwen3			GPT-4.1			GPT-4.1		
Strategy	SLQ-p	SLQ-r	SLQ-F_1	SLQ-p	SLQ-r	SLQ-F_1	SLQ-p	SLQ-r	SLQ-F_1	SLQ-p	SLQ-r	SLQ-F_1
No negatives	66.3	64.0	62.0	57.9	61.5	52.9	**71.6**	**69.6**	67.8	70.7	68.7	67.3
Random	70.8	67.5	65.9	**73.0**	**69.6**	**66.9**	70.3	69.2	**68.4**	**72.2**	**70.6**	**69.6**
NN (concept)	63.1	60.7	58.6	64.8	63.0	58.7	66.2	65.1	64.1	64.7	63.5	61.7
NN (emb)	64.7	61.6	59.7	65.5	63.1	59.0	65.5	64.4	63.4	65.7	63.8	61.8
CONCEPTDEC	**73.4**	**69.2**	**67.2**	70.3	67.5	64.7	62.6	61.6	60.5	68.3	66.5	65.2

5.1 QUERY CONSTRUCTION

Table 1 (a) shows the effect of varying sample count and prompting strategy on SLQ-F_1. Increasing sample count consistently improves performance across all datasets (SynthFace: 48.2% $\rightarrow$ 60.9%; Hancock: 64.3% $\rightarrow$ 68.1%; BigEarthNet: 63.4% $\rightarrow$ 69.7%). However, the optimal prompting strategy varies by dataset. For SynthFace, `TwoGrids` performs best (59.9%), outperforming `Separate` (56.1%). This ranking reverses for Hancock and BigEarthNet, where `Separate` achieves 69.6% and 71.1% respectively, outperforming `TwoGrids` (67.9% and 69.4%). We attribute this to concept structure: SynthFace concepts are localized features comparable across aligned faces, whereas Hancock and BigEarthNet contain heterogeneous images where spatial alignment is less meaningful.

5.2 ACTIVATION THRESHOLD SELECTION

Table 1 (b) examines how the activation threshold q^{lo} affects interpretation quality. For SynthFace, higher thresholds consistently improve SLQ-F_1: the 80th–90th percentile achieves 65.0%–66.4%,

compared to 59.3% with $q^{\text{lo}} = 10$. In contrast, Hancock and BigEarthNet peak at moderate thresholds (40th and 50th percentile: 69.9% and 71.2%), with performance degrading at higher percentiles (68.3% and 67.9% at the 90th). We attribute this to two factors. First, maximally activating samples in real-world data are often outliers: an SAE feature for "Hypochromasia" (reduced tissue staining) may activate most strongly on empty slides—technically low-staining, but not representative of the concept. Second, restricting to top activations reduces sample diversity, limiting the MLLM's ability to identify what unifies positive examples. SynthFace's controlled synthetic distribution avoids such outliers, explaining why higher thresholds help there. The "Extreme" strategy confirms this: it underperforms markedly on Hancock and BigEarthNet (63.6% and 65.0%) where this issue is most pronounced. The strategy from Hoang-Xuan et al. (2024)—finding the M highest activating samples and selecting from them the set of k samples with highest cosine similarity—underperforms quantile-based random selection. Note that the authors only use positives for sample selection, whereas we include negatives in-line with our experimental setup.

5.3 Negative Sample Selection Strategies

Table 2 evaluates negative sample selection strategies (all use samples above the optimal activation threshold from Table 1). For SynthFace, ConceptDecoder achieves the best SLQ-F_1 (67.2%), outperforming random (65.9%) and no negatives (62.0%). However, on Hancock and BigEarthNet, random selection wins (68.4% and 69.6%), while ConceptDecoder underperforms (60.5% and 65.2%). We attribute this to concept structure: SynthFace concepts are often *additive and localized* (glasses, hats can be removed), making counterfactuals well-defined. In contrast, Hancock and BigEarthNet concepts are often *global* ("open water", "dense cellular architecture"), where generating counterfactuals requires replacing entire images. In these datasets, it is also often unclear what the counterfactual of a concept even is (What is the counterfactual of "open water"?). Nearest-neighbour strategies consistently underperform random selection (58.6%–64.1% vs. 65.9%–69.6%). This suggests that truly contrastive real samples—that only differ in the desired concept—are rare. Qualitative examples for SynthFace and Hancock are in Tables I.1 and K.1.

5.4 Rediscovering Labelled Concepts

To validate SLQ metrics, we use them to evaluate supervised SynthFace concepts with ground truth labels. Table D.1 compares generated labels against ground truth. The SLQ-F_1 for labels generated by our automatic interpretation pipeline (85.2%) closely matches that of ground truth labels (88.2%). Random selection achieves SLQ-F_1 of 85.2%, ConceptDecoder 87.5%, and the ground truth labels 88.2%. These results shows that (1) our pipeline successfully rediscovers concepts given sufficiently accurate concept vectors, and (2) the SLQ metrics present a faithful measure of interpretation quality. Perfect scores for ground truth labels are not expected since concept vectors do not perfectly align with true concepts, leading to imperfect activations used as ground truth in SLQ computation.

6 Discussion and conclusion

This paper presents the first systematic investigation of automatic interpretation for visual concepts, introducing SLQ metrics that directly measure label faithfulness via prediction of concept activation on held-out samples.

Our experiments conclude that the optimal interpretation strategy depends on the concept structure. More samples consistently help, but other choices vary: structured datasets with localized, additive concepts (*e.g.* SynthFace) benefit from counterfactual generation and highly activating samples, while natural datasets with global concepts (Hancock, BigEarthNet) perform best using random samples with wider activation ranges, and separate image presentation. Despite this variability, our pipeline achieves strong performance across datasets and autointerp MLLMs, confirming that automatic interpretation of visual concepts is broadly feasible with contemporary models.

Using ConceptDecoder has shown some success, but training dataset-specific models is costly; future work could investigate whether large generative models, such as Qwen-Image (Wu et al., 2025), or FLUX (Labs et al., 2025), can be repurposed to generate counterfactuals directly. We further believe SAE-based concept discovery and MLLM interpretation generalize beyond vision, potentially surfacing patterns humans have not yet identified.

REFERENCES

Ahmed Abdulaal, Hugo Fry, Nina Montaña-Brown, Ayodeji Ijishakin, Jack Gao, Stephanie Hyland, Daniel C Alexander, and Daniel C Castro. An x-ray is worth 15 features: Sparse autoencoders for interpretable radiology report generation. *arXiv preprint arXiv:2410.03334*, 2024.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Computer Vision and Pattern Recognition*, 2017.

Kenza Bouzid, Shruthi Bannur, Felix Meissen, Daniel Coelho de Castro, Anton Schwaighofer, Javier Alvarez-Valle, and Stephanie L. Hyland. Insights into a radiology-specialised multimodal large language model with sparse autoencoders. *arXiv preprint arXiv:2507.12950*, 2025.

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. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

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

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6077–6101. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/bussmann25a.html.

Nick Cammarata, Mark Bissell, Nam Nguyen, Max Loeffler, Eric Ho, Myra Deng, and Daniel Gorton, Livand Balsam. Painting with concepts using diffusion model latents. *Goodfire Research*, 2025. https://www.goodfire.ai/research/painting-with-concepts.

Alba Carballo-Castro, Sonia Laguna, Moritz Vandenhirtz, and Julia E Vogt. Exploiting interpretable capabilities with concept-enhanced diffusion and prototype networks. In *Interpretable AI: Past, Present and Future*, 2024. URL https://openreview.net/forum?id=sBPRqxPrWh.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

Kai Norman Clasen, Leonard Hackel, Tom Burgert, Gencer Sumbul, Begüm Demir, and Volker Markl. reben: Refined bigearthnet dataset for remote sensing image analysis. In *IGARSS 2025-2025 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1264–1268. IEEE, 2025.

Joseph Paul Cohen, Rupert Brooks, Sovann En, Evan Zucker, Anuj Pareek, Matthew P. Lungren, and Akshay Chaudhari. Gifsplanation via latent shift: A simple autoencoder approach to counterfactual generation for chest X-rays. In Mattias Heinrich, Qi Dou, Marleen de Bruijne, Jan Lellmann, Alexander Schläfer, and Floris Ernst (eds.), *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning*, volume 143 of *Proceedings of Machine Learning Research*, pp. 74–104. PMLR, 07–09 Jul 2021. URL https://proceedings.mlr.press/v143/cohen21a.html.

Marion Dörrich, Matthias Balk, Tatjana Heusinger, Sandra Beyer, Hamed Mirbagheri, David J Fischer, Hassan Kanso, Christian Matek, Arndt Hartmann, Heinrich Iro, Markus Eckstein, Antoniu-Oreste Gostian, and Andreas M Kist. A multimodal dataset for precision oncology in head and neck cancer. *Nature Communications*, 16(1):7163, aug 2025.

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. *arXiv preprint arXiv:2406.04093*, 2024.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, et al. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *arXiv preprint arXiv:2410.20526*, 2024.

Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep features. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=NudBMY-tzDr.

Nhat Hoang-Xuan, Minh Vu, and My T. Thai. LLM-assisted concept discovery: Automatically identifying and explaining neuron functions, 2024. URL https://arxiv.org/abs/2406.08572.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pp. 13213–13232. PMLR, 2023.

Neha Kalibhat, Shweta Bhardwaj, Bayan Bruss, Hamed Firooz, Maziar Sanjabi, and Soheil Feizi. Identifying interpretable subspaces in image representations. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Nhat Minh Le, Ciyue Shen, Neel Patel, Chintan Shah, Darpan Sanghavi, Blake Martin, Alfred Eng, Daniel Shenker, Harshith Padigela, Raymond Biju, et al. Learning biologically relevant features in a pathology foundation model using sparse autoencoders. *arXiv preprint arXiv:2407.10785*, 2024.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2. *arXiv preprint arXiv:2408.05147*, 2024.

Krishna Kanth Nakka. Mammo-sae: Interpreting breast cancer concept learning with sparse autoencoders. In *Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care*, pp. 11–20. Springer, 2025.

Thuong Nguyen, Vandana Panwar, Vipul Jamale, Averi Perny, Cecilia Dusek, Qi Cai, Payal Kapur, Gaudenz Danuser, and Satwik Rajaram. Delta marches to autonomously learn histopathology rules by generative latent space traversals. *bioRxiv*, 2025. doi: 10.1101/2025.03.18.643999. URL https://www.biorxiv.org/content/early/2025/07/16/2025.03.18.643999.

Tuomas Oikarinen and Tsui-Wei Weng. Clip-dissect: Automatic description of neuron representations in deep vision networks. *International Conference on Learning Representations*, 2023.

Matthew Lyle Olson, Musashi Hinck, Neale Ratzlaff, Changbai Li, Phillip Howard, Vasudev Lal, and Shao-Yen Tseng. Probing the representational power of sparse autoencoders in vision models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6167–6177, 2025.

OpenAI. Images and vision – learn how to understand or generate images. https://platform.openai.com/docs/guides/images-vision, 2025. Accessed: 2025-11-08.

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

Gonçalo Santos Paulo, Alex Troy Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=EemtbhJOXc.

Sukrut Rao, Sweta Mahajan, Moritz Böhle, and Bernt Schiele. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 444–461, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72980-5.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Samuel Stevens, Wei-Lun Chao, Tanya Berger-Wolf, and Yu Su. Sparse autoencoders for scientifically rigorous interpretation of vision models, 2025. URL https://arxiv.org/abs/2502.06755.

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. URL https://openreview.net/forum?id=MBJJ9Wcpg9.

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. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3681–3691, 2021.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Srikar Yellapragada, Alexandros Graikos, Zilinghan Li, Kostas Triaridis, Varun Belagali, Saarthak Kapse, Tarak Nath Nandi, Ravi K Madduri, Prateek Prasanna, Tahsin Kurc, Rajarsi R. Gupta, Joel Saltz, and Dimitris Samaras. PixCell: A generative foundation model for digital histopathology images, 2025. URL https://arxiv.org/abs/2506.05127.

Éloi Zablocki, Valentin Gerard, Amaia Cardiel, Eric Gaussier, Matthieu Cord, and Eduardo Valle. GIFT: A framework for global interpretable faithful textual explanations of vision classifiers, 2024. URL https://arxiv.org/abs/2411.15605.

Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. Interpreting CLIP with hierarchical sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=5MQQsenQBm.

Andrew Zhang, Guillaume Jaume, Anurag Vaidya, Tong Ding, and Faisal Mahmood. Accelerating data processing and benchmarking of ai models for pathology. *arXiv preprint arXiv:2502.06750*, 2025a.

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

Laura Žigutyte, Tim Lenz, Tianyu Han, Katherine J. Hewitt, Nic G. Reitsam, Sebastian Foersch, Zunamys I. Carrero, Michaela Unger, Alexander T. Pearson, Daniel Truhn, and Jakob Nikolas Kather. Counterfactual diffusion models for mechanistic explainability of artificial intelligence models in pathology. *bioRxiv*, 2024. doi: 10.1101/2024.10.29.620913. URL https://www.biorxiv.org/content/early/2024/11/03/2024.10.29.620913.

Appendix

A Dataset details

SynthFace. SynthFace is a fully synthetic face dataset with photorealistic images rendered from parametric 3D face models (Wood et al., 2021). The procedural generation provides *perfect ground-truth labels* for facial attributes—from accessories (eyeglasses, headphones, hats) to appearance (hair colour, beard, makeup) to context (indoor/outdoor, lighting). We split SynthFace into 10,000 validation, 10,000 test, and 75,801 training samples. From available metadata, we extract binary labels for 14 supervised concepts (Appendix D).

Hancock. Hancock (Dörrich et al., 2025) is a histopathology dataset for precision oncology in head and neck cancer. We reserve 10% of the official train split for validation and use the official *in distribution* test split; slides were processed using Trident (Zhang et al., 2025a) to extract 224×224 patches at 20x magnification. Due to dataset size, we subset to 1% for all steps except unsupervised concept extraction.

BigEarthNet. BigEarthNet (Clasen et al., 2025) consists of Sentinel-2 satellite image patches from 10 European countries, taken between June 2017 and May 2018. We use the official training, validation, and test splits.

The training splits serve as the reference set ($\mathbf{Z}^{\text{ref}}$) for sample selection, activation targets for steering, binarisation thresholds, and concept extraction. All reported results use the test set; validation was used only during development.

B Concept vectors and steering details

Sparse autoencoders. An SAE maps representations to sparse feature vectors via encoder $f(\mathbf{z}) = \sigma(\mathbf{W}^{\text{enc}}\mathbf{z}+\mathbf{b}^{\text{enc}}) \in \mathbb{R}^K$ and reconstructs via decoder $\hat{\mathbf{z}} = \mathbf{W}^{\text{dec}} f(\mathbf{z})+\mathbf{b}^{\text{dec}}$, where σ induces sparsity. Following Bricken et al. (2023), we distinguish between the *detection* and *steering* directions: the encoder weights detect feature presence, while the decoder weights represent the feature's causal effect. For SAE feature i corresponding to concept c, we define:

$$a_c^{\text{enc}}(\mathbf{z}) \coloneqq [f(\mathbf{z})]_i\,, \quad a_c^{\text{dec}}(\mathbf{z}) \coloneqq \mathbf{w}_i^{\text{dec}} \cdot \mathbf{z} + b_i^{\text{dec}}\,, \tag{2}$$

where $\mathbf{w}_i^{\text{dec}}$ and b_i^{dec} are the i-th decoder weight and bias. The encoder activation a_c^{enc} detects feature presence, while a_c^{dec} measures projection onto the causal direction.

Linear probes. For supervised concepts with binary labels $y \in \{0, 1\}$, we train a logistic regression classifier on representations $\mathbf{z} = f_\theta(\mathbf{x})$. The learned weight vector $\mathbf{v}_c \in \mathcal{Z}$ defines the concept direction, and the concept activation is the linear projection:

$$a_c(\mathbf{z}) \coloneqq \langle \mathbf{v}_c, \mathbf{z} \rangle\,. \tag{3}$$

For binarization, we select the threshold τ that maximizes F_1 on the training set: $\text{active}_c(\mathbf{z}) = \mathbb{I}[a_c(\mathbf{z}) > \tau]$. The same concept vector $\mathbf{v}_c$ serves as the steering direction.

Steering equation. Given a sample with representation $\mathbf{z}$, we *steer* along the concept direction $\mathbf{v}_c$ to achieve target activation a^*:

$$\mathbf{z}_c^* = \mathbf{z} + \frac{\mathbf{v}_c}{\|\mathbf{v}_c\|^2}\left(a^* - a_c(\mathbf{z})\right). \tag{4}$$

The normalization by $\|\mathbf{v}_c\|^2$ ensures that the intervention magnitude adapts to concept geometry: $\mathbf{z}_c^*$ achieves exactly a^* regardless of steering vector norm. This makes steering independent of the embedding model, the specific concept or SAE, and the specific sample.

C NEAREST-NEIGHBOUR NEGATIVE SAMPLE SELECTION

The nearest-neighbour strategies aim to find real negative samples that differ from positive samples primarily in the target concept. Starting from positive samples selected using the optimal activation threshold from Table 1 (b), we construct paired negatives through a two-step process: steering followed by retrieval.

Step 1: Steering to target activation. For each positive sample with representation $\mathbf{z}^+$ and activation $a_c(\mathbf{z}^+) > \tau_c$, we steer the representation to a target activation a^*_{neg} in the negative region. To mirror the positive sample selection strategy, we set the target to the center of the corresponding quantile range among negative samples. Specifically, using the optimal q^{lo} from Table 1 (b) with $q^{\text{hi}} = 100$, the target activation is:

$$a^*_{\text{neg}} = \text{quantile}\left(\{a \in A_c^{\text{ref}} : a < \tau_c\}; (q^{\text{lo}} + q^{\text{hi}})/2\right), \tag{5}$$

where $A_c^{\text{ref}} = \{a_c(\mathbf{z}) : \mathbf{z} \in \mathbf{Z}^{\text{ref}}\}$. For example, if positives are selected from the $[q^{\text{lo}}, q^{\text{hi}}] = [90, 100]$ quantile range, we steer to the 95th percentile of the negative activation distribution. Using Equation (4), we obtain the steered representation $\mathbf{z}_c^*$ that achieves exactly this target activation for concept c. The steered representation $\mathbf{z}_c^*$ represents a hypothetical sample that is identical to the original except for the target concept.

Step 2: Nearest-neighbour retrieval. Since $\mathbf{z}_c^*$ may not correspond to any real image, we retrieve the closest real sample from $\mathbf{X}$ as the negative counterpart. We investigate two distance metrics:

- *Embedding space (NN-emb)*: We find the nearest neighbour using cosine distance in the representation space:

$$\mathbf{x}^- = \arg\min_{\mathbf{x} \in \mathbf{X}} \; 1 - \frac{\langle f_\theta(\mathbf{x}), \mathbf{z}_c^* \rangle}{\|f_\theta(\mathbf{x})\| \cdot \|\mathbf{z}_c^*\|}. \tag{6}$$

- *Activation space (NN-concept)*: We find the nearest neighbour using Euclidean distance across all concept activations. Let $\mathbf{a}(\mathbf{z}) = [a_{c_1}(\mathbf{z}), \ldots, a_{c_K}(\mathbf{z})]^\top$ denote the vector of activations for all K concepts. The nearest neighbour is:

$$\mathbf{x}^- = \arg\min_{\mathbf{x} \in \mathbf{X}} \|\mathbf{a}(f_\theta(\mathbf{x})) - \mathbf{a}(\mathbf{z}_c^*)\|_2. \tag{7}$$

D EXTRACTION OF KNOWN CONCEPTS FROM SYNTHFACE

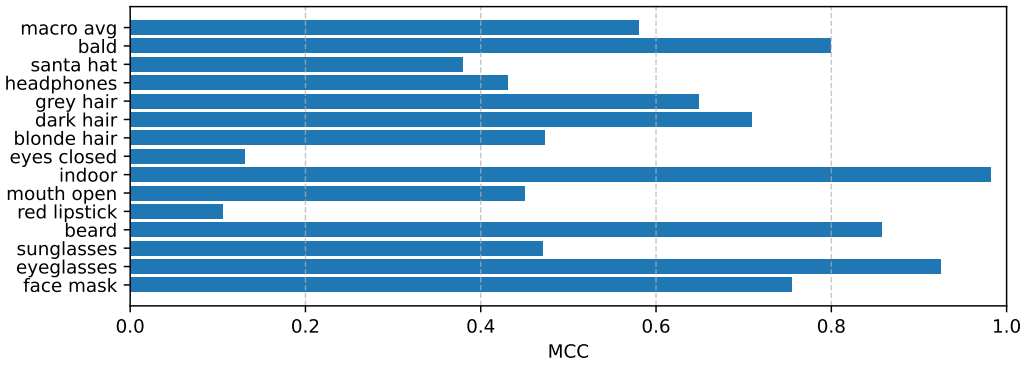


Figure D.1: Per-concept and average Matthews Correlation Coefficient (MCC) for DINOv3 on SynthFace.

In this section, we provide more details on the extraction of concept vectors for the labelled attributes in SynthFace using linear probes. From the SynthFace metadata, we extract binary labels for the following 14 supervised concepts: *face mask*, *eyeglasses*, *sunglasses*, *beard*, *red lipstick*, *mouth open*, *indoor*, *eyes closed*, *blonde hair*, *dark hair*, *grey hair*, *headphones*, *santa hat*, and *bald*. The

parameter space of SynthFace is vast and a single concept may be controlled by multiple parameters. Especially when it comes to continuous parameters (e.g., hair colour), we manually define thresholds to binarize the concepts. This might lead to noisy labels for some concepts. In Figure D.1, we show the classification performance of our linear concept vectors extracted using DINOv3 embeddings.

Using these supervised concept vectors, we validate our interpretation pipeline by checking whether it can rediscover the known ground truth labels. Table D.1 compares the labels produced by the automatic interpretation pipeline against the ground truth for a subset of concepts.

Table D.1: Predicted concept labels for supervised concept rediscovery on SynthFace. We compare the concept labels to those generated by our automatic interpretation pipeline using random real samples (see Table 2, second row)

Ground Truth	Random
"face mask"	"Wearing a face mask"
"eyeglasses"	"Wearing Glasses"
"sunglasses"	"Wearing sunglasses"
"beard"	"Beard or facial hair on chin/jaw"
"red lipstick"	"Dark/Black Lipstick"
"mouth open"	"Wide open mouth (shouting/screaming expression)"
"indoor"	"Indoor Setting"
"eyes closed"	"closed eyes"
"blonde hair"	"Light blonde hair and pale pigmentation"
"dark hair"	"Visible natural hair (no head coverings)"
"grey hair"	"White or gray hair (signs of aging)"
"headphones"	"Wearing a headset with microphone"
"santa hat"	"Christmas/Elf Hat"
"bald"	"Baldness"

E Extraction of unknown concepts with SAEs

We train all SAEs with an expansion factor of 0.5 and k set to 20, resulting in 512 features for the DINO models and CLIP, 576 for SigLIP and SigLIP2, and 768 for UNIv2. We train for 50 epochs [1] with a batch size of 512. The Matryoshka group sizes used were [1/16, 1/8, 1/4, 9/16]. We set the threshold for dead features (used in the auxiliary loss) at 1000.

F ConceptDecoder evaluation

While well aware that a ConceptDecoder is not perfectly reconstructing the input samples, common reconstruction metrics can still provide valuable insights into its performance as a generative model and into its adherence to the conditioning. In Table F.1, we report the Fréchet inception distance (FID), multi-scale structural similarity (MSSSIM), and the cosine similarity between $\mathbf{z}$ and $f_\theta\left(g_\phi\left(\mathbf{z}\right)\right)$ computed using the test sets of SynthFace ($n = 10,000$) and BigEarthNet ($n = 137,367$) and the images generated by a ConceptDecoder conditioned on the respective DINOv3 embeddings.

Table F.1: ConceptDecoder reconstruction metrics.

	SynthFace	BigEarthNet
FID	16.125	25.305
MSSSIM	0.333 ± 0.178	0.317 ± 0.146
Embedding Cosine Similarity	0.746 ± 0.124	0.844 ± 0.102

[1] 3 for Hancock because of the much larger number of samples

G SEMANTIC LABEL QUALITY METRICS

SLQ metrics are obtained using GPT-4.1 in all cases to avoid confounders.

G.1 STATISTICAL SIGNIFICANCE

The evaluation subsampling (selecting at most 100 positive and 100 negative samples per concept) introduces stochasticity into the SLQ metrics. To quantify this variability, we repeated the evaluation 5 times with different random seeds and measured a standard deviation of 0.33 percentage points.

When comparing two methods, the uncertainty in both measurements must be accounted for. Assuming equal variance across methods, the standard error of the difference between two means is $SE_{\text{diff}} = \sqrt{2} \cdot SE$, where $SE = \sigma/\sqrt{n}$. For a two-sided test at significance level $\alpha = 0.05$ with $n = 5$ runs per method, the critical t-value for SynthFace is $t_{0.975,4} \approx 2.78$. The minimum detectable difference (MDD) is therefore:

$$\text{MDD} = t_{\text{crit}} \cdot SE_{\text{diff}} = t_{\text{crit}} \cdot \sqrt{2} \cdot \frac{\sigma}{\sqrt{n}} \approx 0.6 \text{ pp} \tag{8}$$

This means that differences in SynthFace experiments exceeding 0.6 percentage points are statistically significant at $p < 0.05$.

G.2 LIMITATIONS

While SLQ metrics provide a practical means of evaluating the proposed concept labels without ground truth annotations, it is critical to distinguish between *self-consistency* and *correctness*. The SLQ metrics computed in this work essentially measure the agreement between the concept predictor's predictions (i.e. the concept vector or SAE encoder), denoted as $\hat{Y}$, and a set of proxy or heuristics-derived labels, denoted as $\tilde{Y}$ (the proposed labels). However, the ultimate objective is the alignment of $\hat{Y}$ with the latent ground truth, Y. Under conditions where $\tilde{Y}$ and Y are correlated, the SLQ scores can be artificially high even for the wrong label. We quantify this effect below using the SLQ-F_1-score.

Let us define the "confounder strength" c as the symmetric conditional probability between the proposed label $\tilde{Y}$ and the true latent concept Y:

$$P(\tilde{Y} = 1 \mid Y = 1) = P(Y = 1 \mid \tilde{Y} = 1) = c \tag{9}$$

If $\tilde{Y}$ is the "wrong" label for the feature being tracked, c represents the statistical correlation between the proposed misnomer and the true concept label.

Now, consider a predictor $\hat{Y}$ that is tracking the true concept Y with an intrinsic F_1-score of ϕ (where $\phi < 1$). Assuming symmetric precision and recall for algebraic simplicity:

$$P(\hat{Y} = 1 \mid Y = 1) = P(Y = 1 \mid \hat{Y} = 1) = \phi \tag{10}$$

When we evaluate this predictor against the proposed (but incorrect) label $\tilde{Y}$, we are measuring $P(\tilde{Y} \mid \hat{Y})$ and $P(\hat{Y} \mid \tilde{Y})$. Assuming the predictor's errors are independent of the label correlation, we can estimate the measured precision by marginalizing over the ground truth Y:

$$\begin{aligned} \text{Prec}_{\tilde{Y}} &= P(\tilde{Y} = 1 \mid \hat{Y} = 1) && (11)\\ &\approx P(\tilde{Y} = 1 \mid Y = 1) \cdot P(Y = 1 \mid \hat{Y} = 1) && (12)\\ &= c \cdot \phi && (13) \end{aligned}$$

Similarly, the measured recall against the proposed label is:

$$\begin{aligned} \text{Rec}_{\tilde{Y}} &= P(\hat{Y} = 1 \mid \tilde{Y} = 1) && (14)\\ &\approx P(\hat{Y} = 1 \mid Y = 1) \cdot P(Y = 1 \mid \tilde{Y} = 1) && (15)\\ &= \phi \cdot c && (16) \end{aligned}$$

Substituting these into the harmonic mean, the observed SLQ-F_1-score is:

$$\text{SLQ-}F_1 = \frac{2 \cdot (c\phi) \cdot (c\phi)}{c\phi + c\phi} = c \cdot \phi \tag{17}$$

This result ($c \cdot \phi$) implies that the metric is fundamentally ambiguous. A specific SLQ-F_1-score could result from a predictor that moderately learns the proposed label (where $Y \approx \tilde{Y}$), or from a predictor that strongly learns a different concept ($\phi \approx 1$) that happens to be correlated with the proposed label ($c \gg 0$). Thus, while SLQ metrics effectively measure the recoverability of the proposed concept, they cannot inherently rule out the possibility that the model is tracking a correlated confounder.

H MLLM DETAILS

For reproducibility, we set `do_sample=False` for Qwen3-VL and fix temperature and top-p to zero for GPT-4.1.

H.1 TECHNICAL RESTRICTIONS

GPT-4.1. GPT-4.1 can handle up to 500 images with a total payload of 50 MB per request. The token budget per image is 1536 where each token represents a 32×32 patch of an image. Images that exceed the maximum patch size are downsampled internally. We process images with `"detail":"high"`. This means, images are further downsampled if their longest and shortest sides exceed 2048 px and 768 px, respectively. All resizing is performed while preserving the original aspect ratio (OpenAI, 2025).

Qwen3-VL-8B-Instruct. We run Qwen3-VL-8B-Instruct (Yang et al., 2025) locally using the HuggingFace `transformers` library. To ensure reproducibility, we set `do_sample=False`, disabling stochastic sampling. We use the model's default image preprocessing and resolution settings.

H.2 DETAILED PROMPTS

H.2.1 AUTOMATIC INTERPRETATION

Prompt for the TwoGrids strategy.

```
You are an AI researcher.  Your job is to find and explain the
main visual pattern that appears in one set of images but not in
another.

Instructions:

- You will get two groups of images:
1.  Images with the pattern.
2.  Images without the pattern.

- Compare the two groups and describe the single visual concept
that is present in the first group and missing in the second.
- Keep your explanation short and clear.
- Focus on the main concept, not small details.
- Explain why you chose this concept, using examples from both
groups.

Respond in JSON with:
"rationale":  Why you picked this concept, with reference to the
examples.
"concept_label":  A short label for the pattern.
```

Prompt for the OneGrid strategy.

```
You are an AI researcher.  Your job is to find and explain the
main visual pattern that appears in one row of images but not in
another.

Instructions:

- You will get a single grid image with two rows:
1.  The first row (top) contains images with the pattern.
2.  The second row (bottom) contains images without the pattern.

- Compare the two rows and describe the single visual concept that
is present in the first row and missing in the second.
- Keep your explanation short and clear.
- Focus on the main concept, not small details.
- Explain why you chose this concept, using examples from both
rows.

Reply in JSON with:
"rationale":  Why you picked this concept, with reference to the
examples.
"concept_label":  A short label for the pattern.
```

Prompt for the Paired strategy.

```
You are an AI researcher analyzing images to find a single visual
pattern.

You will be shown pairs of images stitched side by side:
1.  The left side always shows images with the pattern.
2.  The right side always shows images without the pattern.

Your task:
- Identify the one visual concept present on the left and missing
on the right.
- Describe this concept clearly and concisely.
- Avoid focusing on minor details; look for the main difference.
- Explain your reasoning, referencing what you see in both sides.

Respond in JSON with:
"rationale":  Why you chose this pattern, based on the images.
"concept_label":  A short label for the visual concept.
```

Prompt for the Separate strategy.

```
You are an AI researcher analyzing images to find a single visual
pattern.

You will be shown pairs of images:
1.  The first image always shows images with the pattern.
2.  The second image always shows images without the pattern.

Your task:
- Identify the one visual concept present on the left and missing
on the right.
- Focus on the most noticeable change between the two images.
- Describe this concept clearly and concisely.
- Avoid focusing on minor details; look for the main difference.
- Explain your reasoning, referencing what you see in both images.
```

```
Respond in JSON with:
"rationale": Why you chose this pattern, based on the images.
"concept_label": A short label for the visual concept.
```

Prompt for the PosOnly strategy.

```
You are an AI researcher analyzing images to find a single visual
pattern.

You will be shown a set of images that contain a common visual
pattern.

Your task:
- Identify the one common visual concept present in the images.
- Describe this concept clearly and concisely.
- Avoid focusing on minor details; look for the main pattern.
- Explain your reasoning, referencing what you see in the images.

Respond in JSON with:
"rationale": Why you chose this pattern, based on the images.
"concept_label": A short label for the visual concept.
```

H.2.2 SEMANTIC LABEL QUALITY

Prompt for the Paired strategy.

```
You are an AI researcher analyzing images.
You will receive:
- A visual concept.
- A series of images.
You must determine the presence of the concept in each image and
give a score of 0 or 1.

Respond with a JSON containing the following keys:
"concept_label": The studied concept.
"scores": A list of dictionaries, each with:
- "image_id": The unique identifier of the image as provided.
- "score": 1 if the concept is present in the image, otherwise 0.

Ensure your response is valid JSON.
```

I AGREEMENT BETWEEN STRATEGIES

We investigate whether selection strategies exhibit complementary strengths within a single dataset. Figure I.1 plots SLQ-F_1 scores of CONCEPTDECODER against random selection for all 512 SynthFace SAE features. The methods show moderate correlation (Pearson $r = 0.45$, $p < 0.001$). Taking the maximum per concept yields 72.1%, compared to 67.2% (CONCEPTDECODER) and 65.9% (random) alone, suggesting partial complementarity. Concepts where both methods achieve high SLQ-F_1 scores ($> 0.75\%$) predominantly relate to facial expressions and facial features, clothing/accessories, hairstyles, and backgrounds, which is encouraging because these are the most prominent varying factors in SynthFace. We have even observed concepts related to lighting conditions.

J GENERALIZATION ACROSS FOUNDATION MODELS

We evaluate concept interpretation on SynthFace using CONCEPTDECODER with CLIP, SigLIP, SigLIP2, and DINOv2 (Table J.1). All models achieve comparable SLQ-F_1 (64.0%–69.2%), with

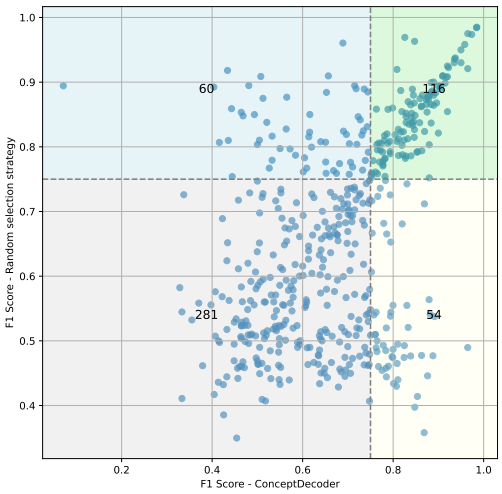


Figure I.1: SLQ-F_1 scores for CONCEPTDECODER vs. random selection across 512 SynthFace SAE features. Dashed lines at 0.75 divide the space into four regions: both methods succeed (116, green), only random succeeds (60, blue), only CONCEPTDECODER succeeds (54, green), neither succeeds (281, grey).

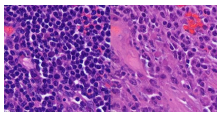
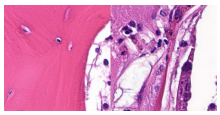
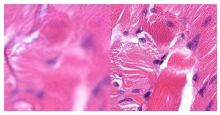
Contrastive pair $(\mathbf{x}_c^+, \mathbf{x}_c^-)$	Predicted label
	"High nuclear density" SLQ-F_1: 0.95
	"Hyaline cartilage matrix" SLQ-F_1: 0.75
	"Blurred tissue architecture" SLQ-F_1: 0.95

Table I.1: Examples from Hancock of interpretable unsupervised concepts. The first two show clinical features; the third is a dataset artefact.

Table J.1: Performance of the automatic interpretation pipeline using counterfactual contrastive pairs generated by CONCEPTDECODER on SynthFace using different foundation models.

Foundation model	SLQ-prec (%)	SLQ-rec (%)	SLQ-F_1 (%)
CLIP	70.9	68.8	67.3
SigLIP	73.5	70.7	69.2
SigLIP2	71.4	68.6	67.0
DINOv2	69.9	66.0	64.0

SigLIP highest, suggesting interpretability reflects genuine representational structure rather than encoder-specific artefacts.

K DETAILED QUALITATIVE RESULTS FOR UNSUPERVISED CONCEPT DISCOVERY ON SYNTHFACE

Table K.2: Detailed results for unsupervised concept discovery on SynthFace. We show the top-performing SAE features interpreted using counterfactual contrastive pairs by a CONCEPTDECODER, with exemplar counterfactual pairs, predicted labels, and SLQ metrics.

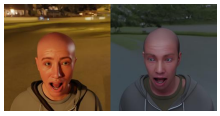
counterfactual pair $(\mathbf{x}_c^+, \mathbf{x}_c^-)$	feature	predicted label	SLQ-F_1	SLQ-prec	SLQ-rec
	7	Eyeglasses	0.985	0.985	0.985
	409	Nighttime outdoor lighting	0.950	0.955	0.950

Table K.2: (cont.).

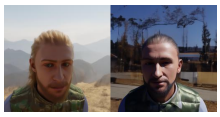
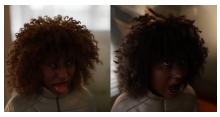
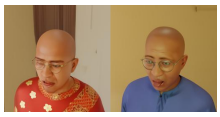
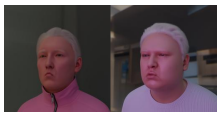
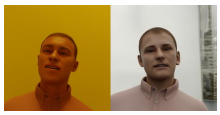
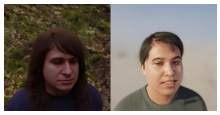
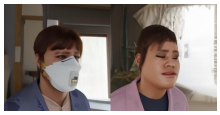
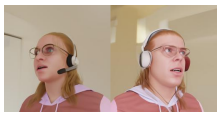
counterfactual pair $(\mathbf{x}_c^+, \mathbf{x}_c^-)$	feature	predicted label	SLQ-F_1	SLQ-prec	SLQ-rec
	159	Defined ringlet curls	0.935	0.936	0.935
	443	Natural Mountainous Outdoor Setting	0.922	0.923	0.922
	266	Tongue sticking out	0.920	0.921	0.920
	265	Red patterned clothing	0.917	0.917	0.917
	3	Zip-up jacket/outerwear	0.910	0.911	0.910
	417	Warm yellow/orange colour cast	0.890	0.893	0.890
	6	Forest Environment Background	0.884	0.899	0.885
	50	Face mask worn	0.884	0.907	0.885
	421	Golden hour backlighting	0.880	0.884	0.880
	223	White hair and beard	0.863	0.882	0.865
	125	Large Full Beard	0.847	0.885	0.850
	150	Topknot/Bun Hairstyle	0.842	0.877	0.845
	439	Short, chunky dreadlocks	0.832	0.857	0.835
	194	Headset with microphone boom	0.817	0.869	0.823

Table K.2: (cont.).

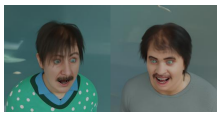
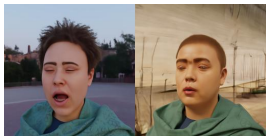
counterfactual pair $(\mathbf{x}_c^+, \mathbf{x}_c^-)$	feature	predicted label	SLQ-F_1	SLQ-prec	SLQ-rec
	152	Long, straight hair	0.812	0.840	0.815
	241	Dramatic overhead shadow lighting	0.804	0.810	0.805
	96	Polka dot patterned clothing	0.793	0.845	0.800

Table K.1: Examples of SAE concepts for SynthFace showing successful interpretations (left) and failure cases (right). We show one exemplar contrastive pair $(\mathbf{x}_c^+, \mathbf{x}_c^-)$ used for automatic interpretation per concept, along with the predicted label and SLQ metrics.

Successful interpretations		Failure cases	
Contrastive pair $(\mathbf{x}_c^+, \mathbf{x}_c^-)$	Predicted label & metrics	Contrastive pair $(\mathbf{x}_c^+, \mathbf{x}_c^-)$	Predicted label & metrics
	"Eyeglasses" SLQ-F_1: 0.98, SLQ-prec: 0.99, SLQ-rec: 0.98		"Old Wooden Interior Background" SLQ-F_1: 0.43, SLQ-prec: 0.76, SLQ-rec: 0.55
	"Wide-brimmed hats" SLQ-F_1: 0.98, SLQ-prec: 0.98, SLQ-rec: 0.98		"Neutral or Negative Facial Expressions" SLQ-F_1: 0.45, SLQ-prec: 0.51, SLQ-rec: 0.51
	"Draped Headscarf" SLQ-F_1: 0.97, SLQ-prec: 0.97, SLQ-rec: 0.97		"Exaggerated facial expressions" SLQ-F_1: 0.55, SLQ-prec: 0.55, SLQ-rec: 0.55