

Efficient and Interpretable Representations: From Medical Representation Learning to Vision-Language Multimodal Representation Engineering

Ali Nasiri-Sarvi

**A Thesis
in
The Department
of
Computer Science and Software Engineering**

**Presented in Partial Fulfillment of the Requirements
for the Degree of
Master of Science (Computer Science) at
Concordia University
Montréal, Québec, Canada**

August 2025

© Ali Nasiri-Sarvi, 2025

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Ali Nasiri-Sarvi**

Entitled: **Efficient and Interpretable Representations: From Medical Representation Learning to Vision-Language Multimodal Representation Engineering**

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

Dr. Yang Wang Chair

Dr. Yang Wang Examiner

Dr. Eugene Belilovsky Examiner

Dr. Mahdi S. Hosseini Supervisor

Dr. Hassan Rivaz Co-supervisor

Approved by

Joey Paquet, Chair
Department of Computer Science and Software Engineering

2025

Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Efficient and Interpretable Representations: From Medical Representation Learning to Vision-Language Multimodal Representation Engineering

Ali Nasiri-Sarvi

Visual representation learning has achieved remarkable progress on natural image benchmarks, but faces critical challenges when deployed in specialized domains like medical imaging. This thesis addresses two interconnected problems: developing efficient architectures that maintain performance while aligning with domain expertise, and creating scalable frameworks for understanding what foundation models learn across different architectures.

We first investigate Vision Mamba architectures for medical applications. For histopathology, we adapt Vision Mamba within the DINO self-supervised learning framework, achieving an 8.21 AUC point improvement over Vision Transformers with comparable parameters on lymph node metastasis detection. Explainability analysis reveals that Vision Mamba focuses on diagnostically relevant cellular features, suggesting better alignment with clinical workflows. For breast ultrasound classification, we demonstrate through transfer learning that Mamba-based architectures achieve statistically significant improvements, with comprehensive analysis showing they are never significantly outperformed by traditional CNN or Vision Transformer baselines.

Our interpretability analysis of pathology foundation models using sparse autoencoders reveals a fundamental scalability problem: each model produces incompatible latent spaces that require separate expert analysis, creating exponential scaling in interpretability effort as foundation models proliferate. To address this limitation, we develop SPARC, a unified framework that enables interpretability analysis across multiple models simultaneously. SPARC introduces a Global TopK mechanism ensuring identical latent dimensions activate across models, and cross-reconstruction

loss enforcing semantic consistency. Our evaluation demonstrates substantial improvements, achieving 84.4% neurons active across all streams compared to 43.6% with traditional approaches, and enabling new capabilities like text-guided spatial attention in vision-only models.

This work contributes efficient architectures for medical applications, identifies fundamental limitations in current interpretability paradigms, and provides a scalable solution that transforms cross-model interpretability from an exponentially scaling manual process into a systematic, unified approach. The results have implications for both medical AI deployment and broader interpretability research as foundation models continue to proliferate across specialized domains.

Acknowledgments

During my two-year Master's program, I have had an enriching and transformative learning experience, made possible by the support and guidance of many remarkable people who have contributed to my academic and personal growth.

I would like to express my deepest gratitude to my supervisors, Mahdi S. Hosseini and Hassan Rivaz, for their excellent guidance and support throughout this program. They provided me with the independence and freedom to develop as a researcher, which has been invaluable to my growth.

I extend my heartfelt thanks to my colleagues and friends: Damien, Cassandre, Amirhossein, Denisha, Vasudev, Sina, Joe, Hojat, Maedeh, and Behnaz. Your camaraderie, intellectual discussions, and collaborative spirit created an inspiring research environment that made even the most challenging days enjoyable.

Je tiens à remercier le Fonds de recherche du Québec – Nature et technologies (FRQNT) pour le soutien financier qui a rendu cette recherche possible et m'a permis de me concentrer sur mes études.

Finally, I want to express my sincere appreciation to my family for their unconditional love, encouragement, and support throughout this academic journey. Your belief in me and understanding during the demanding periods of my Master's program have been a constant source of strength and motivation.

Thank you all for being part of this incredible journey.

Contents

List of Figures	xii
-----------------	-----

List of Tables	xviii
----------------	-------

1 Introduction	1
1.1 Motivation	1
1.1.1 The Evolution of Visual Representation Learning	1
1.1.2 Medical Imaging: A Critical Application Domain	2
1.1.3 The Interpretability Problem	3
1.2 Problem Statement and Research Questions	4
1.2.1 Core Challenges Addressed	4
1.2.2 Research Questions	4
1.3 Methods and Contributions	5
1.4 Publications and Impact	6
1.5 Thesis Organization	7
2 Literature Review	9
2.1 Efficient Representation Learning	10
2.1.1 From Convolutions to Token-Based Models	10
2.1.2 Structured State-Space Models: Linear Complexity with Long-Range Ca- pacity	11
2.1.3 Self-Supervised Pre-Training Enables Scaling	12

2.2	Representation Learning in Medical Domain	14
2.2.1	Representation Learning Challenges in Computational Pathology	14
2.2.2	Multi-Instance Learning in Computational Pathology	15
2.2.3	Foundation Models for Pathology	16
2.2.4	Deep Learning in Breast Ultrasound Imaging	17
2.3	Representation Engineering and Interpretability	18
2.3.1	Classical Computer-Vision Interpretability Methods and Their Limitations	18
2.3.2	Sparse Representations	19
2.4	Aligned Representations	21
2.4.1	Cross-Model Representation Engineering	21
2.4.2	Multimodal Vision-Language Alignment	22
2.5	Concluding Remarks	23
3	Efficient Representation Learning in Medical Domain	25
3.1	Introduction: Why Efficient Representation Learning?	25
3.2	Self-Supervised Vision Mamba for Histopathology Images	27
3.2.1	Methodology	27
3.2.2	Experiments	30
3.2.3	Results	32
3.2.4	Explainability Analysis	35
3.2.5	Discussion	36
3.3	Vision Mamba for Classification of Breast Ultrasound Images	37
3.3.1	Methodology	37
3.3.2	Experimental Setup	39
3.3.3	Results	40
3.3.4	Discussion	42
3.4	Concluding Discussions	43
4	Sparse Autoencoders and interpretability for histopathology images	45
4.1	Sparse Autoencoder Framework for Model Interpretability	46

4.1.1	Mathematical Formulation	46
4.2	SAE Analysis of Pathology Foundation Models	47
4.2.1	The Cross-Model Comparison Challenge	47
4.3	The Scalability Problem in Cross-Model Interpretability	48
4.3.1	Manual Expert Analysis Bottleneck	48
4.3.2	The Need for Unified Interpretability	49
4.4	Toward Cross-Model Interpretability	49
5	Scalable Interpretability and Representation Engineering with SPARC	51
5.1	Introduction: From Problem to Solution	52
5.2	SPARC Method: Shared Sparse Representations Across Models	53
5.2.1	Problem Formulation	53
5.2.2	Architecture Design	54
5.2.3	Training Objective	55
5.3	Experimental Setup	56
5.3.1	Problem Setup and Model Streams	57
5.3.2	Architecture and Training Configuration	57
5.3.3	Datasets and Experimental Scope	58
5.3.4	Baseline Comparisons and Ablation Design	58
5.3.5	Evaluation Framework	59
5.4	Results and Analysis	60
5.4.1	Latent Activation Alignment	60
5.4.2	Quantitative Concept Alignment	62
5.4.3	Monosemantic Concept Recovery	62
5.4.4	Downstream Applications: Semantic Segmentation	63
5.4.5	Downstream Applications: Retrieval Performance	67
5.4.6	Caption \rightarrow Image Retrieval	70
5.4.7	Image \rightarrow Image Retrieval	71
5.4.8	Ablation Analysis	72

5.5	Concluding Remarks	74
5.5.1	Key Achievements	74
5.5.2	Impact on Interpretability Research	75
5.5.3	Future Directions	75
6	Conclusion	77
6.1	Summary of Contributions	77
6.2	Addressing the Core Challenges	78
6.3	Broader Implications	79
6.3.1	For Medical Imaging	79
6.3.2	For Vision and Multimodal AI	80
6.3.3	For Interpretability Research	80
6.4	Limitations and Open Questions	81
6.5	Future Research Directions	81
6.5.1	Short-term Extensions	81
6.5.2	Long-term Vision	82
	Appendix A SPARC Appendix	84
A.1	Experimental Details	84
A.1.1	Datasets	84
A.1.2	Hyperparameters and Training Configuration	84
A.1.3	Probe Implementation Details	86
A.2	Latent Dimension Visualizations	88
A.3	Latent Attribution based on concepts	92
A.3.1	Concept Specific Latents	92
A.3.2	Same image/caption, different latents	97
A.3.3	Using latents of concepts that are not present in the image/caption	97
A.3.4	Limitations of concept-based latent attribution	100
A.4	Cross-Modal Similarity Attribution	102
A.4.1	Cross-modal heatmaps with full captions	102

A.4.2	Same image, different captions	106
A.4.3	Cross-modal attribution limitations	107
A.5	Retrieval Qualitative Results	108
A.5.1	Image → Caption Retrieval (In-Distribution)	108
A.5.2	Caption → Image Retrieval (In-Distribution)	119
A.5.3	Image → Image Retrieval (In-Distribution)	124
A.5.4	External Image → Caption (OOD)	130
A.5.5	Free-Form Caption → Image (OOD)	138
A.5.6	External Image→Image (OOD)	141
Bibliography		145

List of Figures

Figure 3.1	Comparison between different architecture designs. Vim sequential processing allows the model to capture both short-range and long-range dependencies. . . .	28
Figure 3.2	Detailed architecture of VIM within the DINO framework. We modify the Vim model to adapt to input image size for positional embedding interpolation and employ the modified model within DINO as a backbone architecture for self-supervised learning.	28
Figure 3.3	Sequential processing of Vim done on each patch level from slide for feature embedding. This is similar to the lawnmower pattern used for slide navigation by pathologists to study cellular neighbourhoods in the tissue for cancer diagnosis. The information from each patch (i.e. embeddings) are put together to reach to a consensus on the slide level (i.e. aggregation).	30
Figure 3.4	Representative tumor patch with Vim-s heatmap. The red asterisks highlight intracellular mucin in cancer cells. The yellow asterisks highlight stromal features adjacent to cancer cells. (The heatmaps are generated at 10x and overlaid on 40x images.)	35
Figure 3.5	Representative tumor patch with ViT-s heatmap. The red asterisks highlight areas centralized on cancer cells. The yellow asterisks highlight other features, notably a focus of intracellular mucin (top-right) and a stromal cell (middle). (The heatmaps are generated at 10x and overlaid on 40x images.)	36
Figure 3.6	Overview of the Vim model architecture for ultrasound image processing. .	38

Figure 3.7	Overview of the VMamba model architecture with 2D Selective Scan mechanism.	38
Figure 3.8	Abstract comparison between different architecture types for processing spatial relationships in ultrasound images.	39
Figure 4.1	Representative samples from SAE dimension analysis for Phikon and Quilt. We show both strong SAE dimensions with high consistency as well as dimensions with lower consistency in their attributes.	48
Figure 5.1	Detailed architecture of the SPARC model as well as the Global TopK mechanism.	54
Figure 5.2	Top-activating samples for the latent dimension 6463 across three streams (DINO, CLIP-img, CLIP-txt) under four SPARC configurations. Each row shows top-10 images that activate the latent. The CLIP-text stream shows no activations under Local TopK with $\lambda = 1$ due to a dead neuron.	60
Figure 5.3	SPARC enables consistent concept visualization across models and modalities using shared latent dimensions. The figure demonstrates how individual concept-specific latents (bus, cat, balloons) produce coherent spatial heatmaps across DINO and CLIP vision encoders, while also generating meaningful text attribution scores in CLIP’s text encoder when processing full image captions.	64
Figure 5.4	Individual latent attribution using SPARC dimension 279 vs. CLIP similarity baseline. (Above) Saliency maps show the same latent responding to cat-related features across image and text modalities. (Below) Text token relevance scores using SPARC and CLIP text.	66
Figure 5.5	Cross-modal similarity attribution comparing SPARC’s aligned latent space against CLIP similarity baseline. Both methods process the same image and caption, showing different attribution patterns enabled by concept-aligned representations.	66

Figure 5.6	Images retrieved for captions are (1) "In this picture we can see some food products in the glass jars.", (2) "In this image might be taken in the airplane. In this image we can see the speedometers, knobs and some digital screens." Captions are from Open Images test dataset and images are retrieved from the same split. Green boxes indicate when the corresponding image for a caption is successfully retrieved. The second caption shows such a match (Global TopK with CLIP, 3rd rank).	71
Figure 5.7	Cross-model image retrieval results. Each image shows a 4-row layout comparing query stream and reference database combinations. Query image (left) with top-10 retrieved images from reference database (right). All models trained with $\lambda = 1$ on Open Images training set. The retrieval is done on the test set of Open Images. Green border is used to show the exact match of the query image was found.	72
Figure 5.8	Global-vs-Local loss gap showing that Global TopK incurs self-reconstruction costs but provides larger cross-reconstruction benefits.	73
Figure 5.9	Self- and cross-reconstruction NMSE vs. number of latents, demonstrating the trade-offs between latent space capacity and reconstruction quality.	74
Figure A.1	Latent activation examples for dimensions 965, 3667, and 4371 showing top-10 activating images across different configurations.	89
Figure A.2	Additional latent activation examples for dimensions 4950, 5720, 6627, and 8186 showing top-10 activating images across different configurations.	90
Figure A.3	Additional latent activation examples for dimensions 36, 213, 3729, and 6419 showing top-10 activating images across different configurations.	91
Figure A.4	SPARC CLIP text token relevance for kite and balloon concepts. CLIP similarity baseline uses concept names "a kite" and "a balloon" rather than full captions.	93
Figure A.5	SPARC CLIP text token relevance for leopard, tiger, rhinoceros, and red panda concepts. CLIP similarity baseline uses concept names "a leopard", "a tiger", "a rhinoceros", and "a red panda" rather than full captions.	94
Figure A.6	SPARC CLIP text token relevance for banana, croissant, cake, and pasta concepts. CLIP similarity baseline uses concept names "a banana", "a croissant", "a cake", and "pasta" rather than full captions.	95

Figure A.7 SPARC CLIP text token relevance for bicycle, bow and arrow, and wood-burning stove concepts. CLIP similarity baseline uses concept names "a bicycle", "bow and arrow", and "wood-burning stove" rather than full captions.	96
Figure A.8 SPARC CLIP text token relevance for the same image/caption using different concept-specific latent sets. Each panel shows attribution for a different target concept as indicated in the table headers.	98
Figure A.9 SPARC text token relevance for image/captions with a concept that's not present in the sample. Using irrelevant latent dimensions in SPARC causes no gradients. For text scores, all scores are 0.	99
Figure A.10 SPARC CLIP text token relevance for image/captions with a concept that's not present in the sample. SPARC produces non-zero gradients for some of common concepts even in the absence of the concept.	101
Figure A.11 Cross-modal similarity attribution for mixed concepts (cat, apple, butterfly) comparing SPARC's aligned latent space against CLIP similarity baseline.	103
Figure A.12 Cross-modal similarity attribution for animal concepts comparing SPARC's aligned latent space against CLIP similarity baseline.	104
Figure A.13 Cross-modal similarity attribution for object concepts comparing SPARC's aligned latent space against CLIP similarity baseline.	105
Figure A.14 Captions used are "Banana", "Apple", "Kiwi", and "Cat" (non-existent concept).	106
Figure A.15 Captions used are "Cat's Ears", "Cat's Eyes", "Cat's Nose", and "A Cat". We find SPARC fails in the case of detailed heatmaps.	107
Figure A.16 Images retrieved for captions are (1) "In this picture we can see some food products in the glass jars.", (2) "In this image might be taken in the airplane. In this image we can see the speedometers, knobs and some digital screens." Captions are from Open Images test dataset and images are retrieved from the same split. Green boxes indicate when the corresponding image for a caption is successfully retrieved. The second caption shows such a match (Global TopK with CLIP, 3rd rank).	120

Figure A.17 Images retrieved for captions are (1) "In this image in the center there is one bird flying, and in the background there is sky.", (2) "In this image we can see a table on which some glasses are there in which some food items and straws are there and we can see a pot like structure. On the left side we can see a person hand. In the background we can see some posters and bottles.", (3) "In the picture we can see an ice cream with a green and brown color cream.", and (4) "In this image we can see a wooden basket placed on the ground, we can also see the photo frame on the wall." Captions are from Open Images test dataset and images are retrieved from the same split. Green boxes indicate when the corresponding image for a caption is successfully retrieved. The third caption shows such a match (Global TopK with CLIP, 1st rank).	121
Figure A.18 Images retrieved for captions are (1) "A woman stands in the dining area at the table.", (2) "A shower curtain sits open in an empty and clean bathroom.", (3) "a man in a blue shirt and red tie.", and (4) "These people are going to have pizza and wine." Captions are from Open Images test dataset and images are retrieved from the same split. Green boxes indicate when the corresponding image for a caption is successfully retrieved.	122
Figure A.19 Images retrieved for captions are (1) "White and orange fur lays on a white blanket.", (2) "A kitchen that has carpeted floors and wooden cabinets.", (3) "A street scene with a horse pulling a white carriage.", and (4) "A large dogs comfortably sleeping on someones bed". Captions are from Open Images test dataset and images are retrieved from the same split. Green boxes indicate when the corresponding image for a caption is successfully retrieved.	123
Figure A.20 Cross-model image retrieval results. Each image shows a 4-row layout comparing query stream and reference database combinations. Query image (left) with top-10 retrieved images from reference database (right). All models trained with $\lambda = 1$ on Open Images training set. The retrieval is done on the test set of Open Images. Green border is used to show the exact match of the query image was found.	125

Figure A.21 Cross-model image retrieval results. Each image shows a 4-row layout comparing query stream and reference database combinations. Query image (left) with top-10 retrieved images from reference database (right). All models trained with $\lambda = 1$ on Open Images training set. The retrieval is done on the test set of Open Images. Green border is used to show the exact match of the query image was found. 126

Figure A.22 Cross-model image retrieval results. Each image shows a 4-row layout comparing query stream and reference database combinations. Query image (left) with top-10 retrieved images from reference database (right). All models trained with $\lambda = 1$ on Open Images training set. The retrieval is done on the test set of Open Images. Green border is used to show the exact match of the query image was found. 127

Figure A.23 Cross-model image retrieval results. Each image shows a 4-row layout comparing query stream and reference database combinations. Query image (left) with top-10 retrieved images from reference database (right). All models trained with $\lambda = 1$ on MS COCO training set. The retrieval is done on the validation set of MS COCO. Green border is used to show the exact match of the query image was found. 129

Figure A.24 Images retrieved for captions: (1) "Black cat", (2) "Orange cat", (3) "A child flying a red kite on a sunny beach", (4) "A barista making coffee behind a counter." The images are retrieved from the test set of the Open Images dataset, but the captions are not part of the dataset. 138

Figure A.25 Images retrieved for captions: (1) "People crossing a busy city street in the rain", (2) "A passenger airplane flying in a clear blue sky", (3) "A dog running through a grassy field", (4) "A man reading a newspaper at a bus stop." The images are retrieved from the test set of the Open Images dataset, but the captions are not part of the dataset. 139

Figure A.26 Images retrieved for captions: (1) "A person riding a bicycle on a country road", (2) "A bowl of fresh strawberries on a kitchen counter", (3) "A train arriving at an underground metro station", (4) "Books stacked on a wooden desk." The images are retrieved from the test set of the Open Images dataset, but the captions are not part of the dataset. 140

Figure A.27 Cross-model image retrieval results using OOD query images. The refer- ences are from the Open Images test set.	142
Figure A.28 Cross-model image retrieval results using OOD query images. The refer- ences are from the Open Images test set.	143
Figure A.29 Cross-model image retrieval results using OOD query images. The refer- ences are from the Open Images test set.	144

List of Tables

Table 3.1	Camelyon16 dataset distribution	30
Table 3.2	Pre-training patch distribution (tumor patches come from tumor slides without ROI filtering)	31
Table 3.3	PCam patch-classification dataset (balanced classes: counts shown as $2 \times N$)	31
Table 3.4	Trainable parameters (rounded to the nearest million)	32
Table 3.5	Slide-level classification on Camelyon16. AUC is the primary comparison metric.	33
Table 3.6	Linear-evaluation accuracy on PCam-224 datasets.	33
Table 3.7	Effect of zooming level on MIL.	34
Table 3.8	Effect of zooming level on patch classification (linear evaluation).	34
Table 3.9	Transfer learning results for BUSI+B dataset. AUC and accuracy values scaled 0-100, averaged over five runs.	41
Table 3.10	Transfer learning results for BUSI dataset. AUC and accuracy values scaled 0-100, averaged over five runs.	41
Table 3.11	Transfer learning results for B dataset. AUC and accuracy values scaled 0- 100, averaged over five runs.	42
Table 5.1	Neuron activation patterns and stream-specific dead neuron rates across 8192 latent dimensions. Mixed patterns indicate partial cross-stream alignment where only 1/3 or 2/3 of the streams are active for the same latent. CI = CLIP-image, CT = CLIP-text, D = DINO.	61

Table 5.2 Mean Jaccard similarity across Open Images taxonomy depths, grouped by TopK type and cross-loss weight λ . Depth 0 corresponds to full collapse into the root category (Entity), while depth 5 corresponds to no collapse (leaf-level granularity).	62
---	----

Table 5.3 Mean probe loss (lower is better) across 432 Open Images binary classification tasks.	63
---	----

Table 5.4 Weakly supervised segmentation results on MS COCO comparing cross-modal similarity approaches. All methods use Chefer, Gur, and Wolf (2021b) relevance map generation. SPARC methods compute similarities in the concept-aligned sparse latent space, while baselines use standard feature spaces.	65
--	----

Table 5.5 Latent alignment R@1 scores across datasets and training regimes. CI = CLIP_IMG, CT = CLIP_TXT, D = DINO.	67
---	----


Table 5.6 The query image  is used to retrieve the captions.	
Green color is for the original caption from the dataset.	69


Table 5.7 The query image  is used to retrieve the captions. Green color is for the original caption from the dataset.	76
--	----




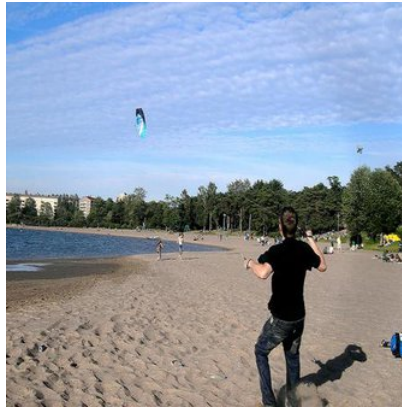
Table A.1	 <p>The query image is used to retrieve the captions. None of the retrieved text is the exact caption of the query image, but still highly relevant captions.</p>	109
Table A.2	 <p>The query image is used to retrieve the captions.</p>	110
Table A.3	 <p>The query image is used to retrieve the captions.</p>	111
Table A.4	 <p>The query image is used to retrieve the captions.</p>	112
Table A.5	 <p>The query image is used to retrieve the captions.</p>	113

Table A.6 The query image is used to retrieve the



captions. Green color is for the original caption from the dataset. 114

Table A.7 The query image is used to retrieve the



captions. Green color is for the original caption from the dataset. 115

Table A.8 The query image is used to retrieve the



captions. Green color is for the original caption from the dataset. 116

Table A.9 The query image is used to retrieve the



captions. Green color is for the original caption from the dataset. 117

Table A.10 The query image is used to retrieve the



captions. Green color is for the original caption from the dataset. 118

Table A.11 The query image is used to retrieve the



captions. 131

Table A.12 The query image is used to retrieve the



captions. 132

Table A.13 The query image		is used to retrieve the	
captions.			133

Table A.14 The query image		is used to retrieve the	
captions.			134


Table A.15 The query image		is used to retrieve the	
captions.			135


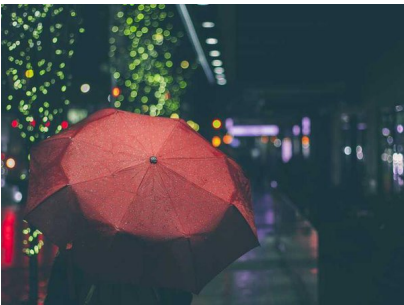
Table A.16 The query image		is used to retrieve the	
captions.			136

Table A.17 The query image		is used to retrieve the	
captions.			137

Chapter 1

Introduction

1.1 Motivation

1.1.1 The Evolution of Visual Representation Learning

Visual representation learning has evolved rapidly from handcrafted features to deep learned representations that now surpass human-level performance on many benchmarks. This transformation began with AlexNet’s breakthrough on ImageNet ([Deng et al., 2009](#); [Krizhevsky, Sutskever, & Hinton, 2012](#)), followed by increasingly sophisticated architectures from VGG ([Simonyan & Zisserman, 2015](#)) and ResNet ([He, Zhang, Ren, & Sun, 2016](#)) to Vision Transformers ([Dosovitskiy et al., 2021](#)) and modern foundation models.

However, a notable gap has emerged between research achievements and practical deployment. While models achieve impressive performance on natural image benchmarks, their translation to specialized domains with critical safety and interpretability requirements faces fundamental challenges. The computational demands of state-of-the-art architectures, combined with their black-box decision-making processes, create barriers to adoption in resource-constrained environments where trust and explainability are top priorities.

This gap becomes most apparent in medical domain, where model failures have consequences far beyond benchmark metrics. Healthcare simultaneously demands cutting-edge performance,

computational efficiency for point-of-care deployment, and interpretability for clinical trust: requirements that existing approaches struggle to satisfy together. As foundation models evolve across specialized domains, bridging this gap between research capabilities and practical deployment needs becomes increasingly critical for sustainable AI advancement.

1.1.2 Medical Imaging: A Critical Application Domain

Medical imaging presents unique computational challenges that expose fundamental limitations in current AI paradigms. In computational pathology, whole-slide images routinely exceed $100,000 \times 100,000$ pixels ([Hosseini et al., 2024](#)), making them several orders of magnitude larger than natural images and forcing patch-based processing that disrupts global context. These gigapixel images must be analyzed under weak supervision constraints, as slide-level diagnostic labels are often the only available signals, while detailed annotations from pathologists remain time-consuming and expensive to obtain ([Campanella et al., 2019](#)).

Domain shifts further complicate deployment, as institutional differences in staining protocols, scanners, and patient populations create significant performance drops during external validation. Studies have shown models can lose 15-25 F_1 points when applied to unseen devices ([Aubreville et al., 2023](#)), with some cases approaching 30-point degradations. Similarly, breast ultrasound imaging faces low signal-to-noise ratios, heterogeneous tissue textures, and modality-specific artifacts that demand robust architectures capable of operating on resource-constrained point-of-care devices ([Afrin, Larson, Fatemi, & Alizad, 2023](#)).

These domain-specific constraints highlight a critical gap between AI research advances and clinical adoption. While Vision Transformers achieve impressive results on ImageNet, their quadratic computational complexity becomes prohibitive for gigapixel pathology images or real-time ultrasound analysis. More fundamentally, the black-box nature of modern architectures conflicts with clinical requirements for interpretable, trustworthy decision support systems where understanding model reasoning is essential for diagnostic confidence and regulatory approval.

1.1.3 The Interpretability Problem

The rapid development of foundation models has created a fundamental scalability problem in interpretability research. While performance benchmarking scales linearly with the number of models (each model evaluated once to obtain comparable metrics), interpretability analysis scales exponentially, requiring separate examination of each architecture’s learned representations. Traditional approaches like gradient-based attribution ([Selvaraju et al., 2017](#); [Simonyan, Vedaldi, & Zisserman, 2013](#)) and sparse coding methods ([Bricken et al., 2023](#)) produce model-specific insights that cannot be systematically compared across architectures, forcing researchers to analyze each new model individually.

This limitation becomes particularly pronounced in specialized domains like computational pathology, where foundation models such as Phikon ([Filiot et al., 2023](#)), UNI ([R. J. Chen et al., 2024](#)), Virchow ([Vorontsov et al., 2023](#)), and CHIEF ([X. Wang et al., 2024](#)) are being developed at an accelerating pace. Understanding what concepts these models have learned requires expert pathologist validation for each architecture separately, creating an unsustainable bottleneck that cannot keep pace with model development. The resulting interpretability debt grows exponentially as more sophisticated models emerge.

In medical domains, this interpretability problem has significant implications for clinical adoption and trust. Healthcare providers need a systematic understanding of model behavior across different architectures to make informed deployment decisions, yet current interpretability paradigms offer no framework for comparing what different models learn or identifying shared versus distinct concept representations. As AI systems move into high-stakes clinical applications, the inability to systematically understand and compare model representations becomes a fundamental barrier to responsible deployment and regulatory approval.

1.2 Problem Statement and Research Questions

1.2.1 Core Challenges Addressed

This thesis tackles three specific problems that emerged from our work in medical imaging and interpretability research. In medical imaging, we investigated whether Vision Mamba’s (Y. Liu et al., 2024; L. Zhu et al., 2024) sequential processing could offer advantages over Vision Transformers (Dosovitskiy et al., 2021), particularly given the alignment between Mamba’s scanning patterns and how pathologists examine tissue slides. While Vision Transformers achieve strong performance, their quadratic attention complexity and the way they process spatial information may not be optimal for medical imaging workflows where understanding spatial relationships and cellular neighborhoods is critical (Molin, Fjeld, Mello-Thoms, & Lundström, 2015).

In interpretability research, we discovered a fundamental bottleneck: analyzing what foundation models learn requires separate examination of each model’s representations. When we applied sparse autoencoders (Bricken et al., 2023) to pathology models like Phikon (Filiot et al., 2023) and Quilt (Ikezogwo et al., 2023), each produced incompatible latent spaces with no systematic way to compare concepts across models. This creates an exponential scaling problem as more foundation models are developed in computational pathology, including UNI (R. J. Chen et al., 2024), Virchow (Vorontsov et al., 2023), and CHIEF (X. Wang et al., 2024).

The third challenge emerged from this limitation: we need methods to compare and align representations across different models simultaneously. Current interpretability tools force researchers to analyze each model individually, making it impossible to systematically understand how different architectures, training objectives, or domain adaptations affect what models learn. This prevents the kind of systematic model comparison that would inform better architectural choices and clinical deployment decisions.

1.2.2 Research Questions

Our research addresses three specific questions.

(1) Do Vision Mamba architectures offer advantages for medical imaging tasks, and does their

sequential processing align with clinical expertise? This question is critical for the affordable deployment of AI models in clinical settings. Mamba offers memory advantages during inference compared to Transformer counterparts, while still being able to capture long-range dependencies. If Mamba-based models are proven effective in the medical domain, they could open new possibilities for model deployment. Moreover, clinician trust requires models whose reasoning aligns with expert workflows—without this alignment, even high-performing models may face adoption resistance. Mamba’s sequential processing could align with expert reasoning in certain domains, such as pathology.

(2) What can we learn from applying interpretability tools to pathology foundation models, and where do current methods break down? Understanding this is essential because medical AI interpretability directly impacts patient safety, regulatory approval, and clinical trust. While many existing interpretability techniques, such as saliency maps or attention visualizations, offer some insights into model behavior, they often fall short in truly unraveling the model’s representations. Their faithfulness is often questionable, as these methods may highlight correlations rather than provide clinically meaningful or causally grounded explanations, which limits their reliability in medical applications.

(3) Can we develop frameworks that enable interpretability analysis across multiple models simultaneously, eliminating the need for separate per-model examination? This capability is crucial because the current paradigm of analyzing each model individually cannot keep pace with AI development. Without unified interpretability, we’re forced to choose between deploying models we don’t understand or accepting an ever-growing interpretability debt that limits responsible AI advancement across all domains, not just medicine.

1.3 Methods and Contributions

We addressed the medical imaging challenges by adapting Vision Mamba architectures for domain-specific applications. For histopathology, we modified Vision Mamba ([L. Zhu et al., 2024](#)) to work within the DINO self-supervised learning framework ([Caron et al., 2021](#)), implementing positional encoding interpolation to handle varying image sizes required by DINO’s global-local view

training. This adaptation achieved an 8.21 AUC point improvement over Vision Transformers with comparable parameters on Camelyon16 ([Bejnordi et al., 2017](#)), with explainability analysis revealing that Vision Mamba focuses on diagnostically relevant cellular features like intracellular mucin. For breast ultrasound classification, we demonstrated through transfer learning that VMamba ([Y. Liu et al., 2024](#)) achieves statistically significant improvements, with comprehensive analysis showing Mamba-based models were never significantly outperformed by traditional architectures.

Our interpretability investigation using sparse autoencoders ([Bricken et al., 2023](#)) on pathology foundation models like Phikon ([Filiot et al., 2023](#)) and Quilt ([Ikezogwo et al., 2023](#)) revealed a fundamental scalability problem: each model produces incompatible latent spaces with no systematic way to compare concepts across architectures. This limitation means that interpretability analysis effort grows exponentially with the number of models, creating a bottleneck as foundation models proliferate in computational pathology and other specialized domains.

To address this problem, we developed SPARC (**S**parse Autoencoders for **A**ligned **R**epresentation of **C**oncepts), a unified framework that enables interpretability analysis across multiple models simultaneously. SPARC introduces two key innovations: (1) a Global TopK mechanism that ensures identical latent dimensions activate across different models for the same input, and (2) cross-reconstruction loss that enforces semantic consistency by requiring each model’s representation to reconstruct features from other models. Our evaluation demonstrates improvements in concept alignment, achieving 84.4% neurons active across all streams compared to 43.6% with traditional approaches, and enabling new capabilities like text-guided spatial attention in vision-only models.

1.4 Publications and Impact

The work presented in this thesis has resulted in the following peer-reviewed publications:

- **Ali Nasiri-Sarvi**, Vincent Quoc-Huy Trinh, Hassan Rivaz, Mahdi S. Hosseini. "Vim4Path: Self-Supervised Vision Mamba for Histopathology Images." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), 9th CVMI (Computer Vision for Microscopy Image Analysis) Workshop, pp. 6894-6903, 2024. (Oral presentation) [\[Paper\]](#) [\[Code\]](#)

- **Ali Nasiri-Sarvi**, Mahdi S. Hosseini, Hassan Rivaz. "Vision Mamba for Classification of Breast Ultrasound Images." In Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care, International Conference on Medical Image Computing and Computer-Assisted Intervention (**MICCAI**), pp. 148-158. Springer, 2024. (Oral presentation) [\[Paper\]](#) [\[Code\]](#)

Additional work on scalable cross-model interpretability frameworks (Chapter 5) is available as a preprint:

- **Ali Nasiri-Sarvi**, Hassan Rivaz, Mahdi S. Hosseini. "SPARC: Concept-Aligned Sparse Autoencoders for Cross-Model and Cross-Modal Interpretability." *arXiv preprint arXiv:2507.06265*, 2025. [\[arXiv\]](#) [\[Code\]](#)

1.5 Thesis Organization

This thesis is organized as follows. Chapter 2 provides a comprehensive literature review covering efficient representation learning, medical domain applications, interpretability methods, and aligned representations across models. The review establishes the theoretical foundation and identifies gaps that motivate our research contributions.

Chapter 3 presents our work on efficient representation learning for medical imaging, including Vim4Path for histopathology and Vision Mamba for breast ultrasound classification. We demonstrate how Vision Mamba architectures can be adapted for medical domains and provide evidence that sequential processing aligns with clinical workflows while achieving superior performance.

Chapter 4 analyzes pathology foundation models using sparse autoencoders, revealing fundamental limitations in current interpretability approaches. We demonstrate the exponential scaling problem that emerges when trying to understand and compare representations across multiple foundation models, motivating the need for unified interpretability frameworks.

Chapter 5 introduces SPARC, our solution to the cross-model interpretability challenges identified in the previous chapter. We present the Global TopK mechanism and cross-reconstruction loss innovations that enable systematic concept comparison across heterogeneous architectures, transforming interpretability analysis from an exponentially scaling problem into a scalable approach.

Chapter 6 summarizes our findings, discusses broader implications for visual representation learning and medical AI, and outlines future research directions that build upon the foundations established in this work.

Chapter 2

Literature Review

Visual representation learning has evolved from handcrafted features to foundation models, driven by computational efficiency and representational power demands. As visual data grows exponentially across domains like medical imaging, learning meaningful, transferable representations becomes increasingly critical. This literature review examines key developments in representation learning through five interconnected themes that illustrate the progression from basic feature extraction to sophisticated multimodal models.

Section 2.1 discusses architectural evolution from convolutional neural networks through attention-based transformers to structured state-space models, examining how each paradigm addressed predecessor limitations while introducing new scale and efficiency capabilities. Self-supervised learning emerged as a unifying framework, enabling architectures to leverage vast unlabeled datasets.

Section 2.2 addresses unique challenges in medical domains, where gigapixel resolutions, weak supervision, and domain shifts demand specialized approaches. We examine how multi-instance learning frameworks, foundation models, and transfer learning strategies address constraints in computational pathology and ultrasound imaging.

Section 2.3 explores the interpretability of learned representations. We review gradient-based and perturbation-based methods alongside recent developments in sparse coding and mechanistic interpretability, highlighting the tradeoff between model expressiveness and human understanding.

Section 2.4 examines aligned representations across different models and modalities. As the

field moves toward more general AI systems, the ability to compare, combine, and transfer representations becomes crucial for building robust, interpretable models operating across different modalities.

Section 2.5 discusses these developments to identify current limitations, emerging opportunities, and open questions motivating subsequent research contributions. We emphasize not only achievements but fundamental challenges in bridging computational efficiency, interpretability, and real-world applicability in specialized domains.

2.1 Efficient Representation Learning

Building efficient yet expressive visual representations has required multiple successive architectural and training innovations, each reducing constraints that previously limited performance, scale, and data efficiency. In this section, we examine how these foundational developments, including convolutional networks, transformers, state-space models, and self-supervised paradigms, have shaped the modern landscape of efficient representation learning.

2.1.1 From Convolutions to Token-Based Models

Early convolutional neural networks (CNNs) trace back to **LeNet-5** (LeCun, Bottou, Bengio, & Haffner, 2002), which introduced local receptive fields, weight sharing, and back-propagation to recognize handwritten digits. Two decades later, **AlexNet** (Krizhevsky et al., 2012) leveraged GPUs, ReLU activations, and large-scale data augmentation to scale CNNs to ImageNet (Deng et al., 2009) and ignite the modern deep learning era. **VGG** (Simonyan & Zisserman, 2015) demonstrated that deeper hierarchies built from uniform 3×3 convolutions further improve accuracy, while **ResNet**’s (He et al., 2016) residual shortcuts enabled hundreds of layers to train stably. Subsequent architectures targeted efficiency: **Inception** (Szegedy et al., 2015) mixed multiple receptive-field sizes within a block, **DenseNet** (G. Huang, Liu, Van Der Maaten, & Weinberger, 2017) maximized feature reuse via dense connectivity, the **MobileNet** family-*v1* (A. G. Howard et al., 2017) with depth-wise separable convolutions, *v2* (Sandler, Howard, Zhu, Zhmoginov, & Chen, 2018) introducing inverted residuals and linear bottlenecks, and *v3* (A. Howard et al., 2019) combining neural

architecture search with squeeze-and-excite modules-enabled deployment on edge devices, and **EfficientNet** (M. Tan & Le, 2019) formalized compound scaling to optimize parameters, FLOPs, and accuracy jointly.

Self-attention reshaped sequence modeling in NLP: the attention mechanism in (Bahdanau, Cho, & Bengio, 2015) enabled dynamic focus within encoder–decoder translation, and the **Transformer** architecture (Vaswani et al., 2017) replaced recurrence entirely with self-attention to unlock massive parallelism.” Adapting these ideas to vision, the **Vision Transformer (ViT)** (Dosovitskiy et al., 2021) partitioned images into patch tokens and, with large-scale pre-training, surpassed CNN baselines on ImageNet. Follow-up works addressed data hunger and locality. **DeiT** (Touvron et al., 2021) employed knowledge distillation and heavy augmentation to train ViTs from scratch on ImageNet, while the **Swin Transformer** (Z. Liu et al., 2021) introduced shifted, hierarchical windows that combine local inductive bias with linear complexity. Different variants of ViT then pushed specific frontiers: **Pyramid ViT (PVT)** (W. Wang et al., 2021) built feature pyramids for dense prediction, **Convolutional ViT (CvT)** (Wu et al., 2021) fused depthwise convolutions with attention for early token embedding, **Tokens-to-Token ViT (T2T-ViT)** (Yuan et al., 2021) iteratively restructured tokens to model local structure before global attention, and **CrossViT** (C.-F. R. Chen, Fan, & Panda, 2021) employed parallel branches with different token sizes to capture multi-scale information.

2.1.2 Structured State-Space Models: Linear Complexity with Long-Range Capacity

Although Transformers capture global context, their $\mathcal{O}(L^2)$ attention cost caused a search for linear-time alternatives as vanilla RNNs suffer from limited context length and vanishing gradients. The **HiPPO** framework (Gu, Dao, Ermon, Rudra, & Ré, 2020) formalized continual memory by projecting the recent input history onto orthogonal polynomial bases, yielding an online update rule that scales as $\mathcal{O}(N)$ per step, where N is the approximation order. Building on this foundation, **Linear State-Space Layers (LSSL)** (Gu et al., 2021) parameterize learnable continuous-time state equations that can be efficiently computed via convolution, unifying RNNs, CNNs, and ordinary differential equations in a single module. However, LSSLs suffered from prohibitive computational

costs, requiring $\mathcal{O}(N^2L)$ operations and $\mathcal{O}(NL)$ space to compute the convolution kernel during training, where N is the state dimension and L is the sequence length. **S4** (Gu, Goel, & Ré, 2022) formalized Structured State Space Models (SSMs) and addressed this bottleneck by parameterizing the state transition matrix using a Normal Plus Low-Rank (NPLR) representation, decomposing it as the sum of a diagonalizable normal matrix and a low-rank correction, reducing the kernel computation complexity to $\tilde{\mathcal{O}}(N + L)$ time and $\mathcal{O}(N + L)$ memory. While S4 maintained linear time-invariant (LTI) dynamics, **Mamba** (Gu & Dao, 2024) introduced *selective* state updates: an input-dependent selection mechanism modulates the input matrix B , output matrix C , and discretization parameter Δ at each timestep, while a fused CUDA kernel exploits diagonal structure to run $4\text{--}5\times$ faster than Flash Attention on long-range language and audio tasks.

Adapting SSMs to images required bridging 1-D sequences and 2-D grids. **Vision Mamba (Vim)** (L. Zhu et al., 2024) flattens patches, injects 2-D positional encodings, and applies bidirectional SSM blocks along both raster directions; on ImageNet-1k (Deng et al., 2009) it matches DEiT-S (Touvron et al., 2021) accuracy with $3\times$ fewer FLOPs and 86 % less memory. **VMamba** (Y. Liu et al., 2024) introduces a *Selective Scan 2-D* operator that sweeps rows and columns independently but shares parameters across directions, achieving SWIN-T-level (Z. Liu et al., 2021) performance on classification while also excelling at detection (COCO (Lin et al., 2014) mAP 47.2) and segmentation (ADE20k (B. Zhou et al., 2017) mIoU 48.1) under linear complexity. Collectively, these works position SSMs as a hardware-efficient alternative to attention, particularly well-suited for memory-constrained and long-sequence settings.

2.1.3 Self-Supervised Pre-Training Enables Scaling

While efficient architectures define the computational backbone of visual models, their success increasingly depends on how well they are pretrained. Self-supervised learning (SSL) emerged as a powerful alternative to supervised labels by designing *pretext tasks* that encourage models to learn generalizable structure from raw data. Predicting the relative position of image patches (Doersch, Gupta, & Efros, 2015), re-assembling jigsaw puzzles (Noroozi & Favaro, 2016), estimating camera egomotion (Agrawal, Carreira, & Malik, 2015), recoloring grayscale images (R. Zhang, Isola, & Efros, 2016), and recognizing rotated images (Gidaris, Singh, & Komodakis, 2018) all demonstrated

that surrogate objectives could yield transferable features without human labels.

Clustering-based approaches scaled these ideas to larger datasets. **DeepCluster** (Caron, Bojanowski, Joulin, & Douze, 2018) alternated k -means assignments with network updates; **SeLa** (Asano, Rupprecht, & Vedaldi, 2020) formulated self-labeling as an optimal-transport problem; and **SwAV** (Caron et al., 2020) replaced pairwise comparisons with online “swapped” assignments, bringing ResNet-50 within 1–2% top-1 of supervised ImageNet.

Contrastive methods reshaped SSL by showing that instance discrimination with strong augmentation can yield high-quality features. **SimCLR** (T. Chen, Kornblith, Norouzi, & Hinton, 2020) maximizes agreement between two augmented views of an image using a contrastive loss and a projection head. **SimCLR-v2** (T. Chen, Kornblith, Swersky, Norouzi, & Hinton, 2020) demonstrated that scaling model depth and training duration further improves representation quality. **MoCo** (He, Fan, Wu, Xie, & Girshick, 2020) introduced a momentum-updated key encoder and a queue of negatives to remove the need for large batches. **MoCo-v2** (X. Chen, Fan, Girshick, & He, 2020) added stronger augmentations and an MLP head to boost performance. **MoCo-v3** (X. Chen, Xie, & He, 2021) adapted the momentum-contrast framework to Vision Transformers, confirming its effectiveness beyond CNN backbones.

Following the development of contrastive frameworks, a series of methods emerged that removed the need for explicit negative pairs. **BYOL** (Grill et al., 2020) introduced a target network updated via exponential moving average and trained the online encoder to predict its outputs, achieving strong performance without contrastive loss. **SimSiam** (X. Chen & He, 2021) showed that a stop-gradient on one branch is sufficient to prevent collapse, simplifying the architecture by removing both momentum encoders and negatives. **Barlow Twins** (Zbontar, Jing, Misra, LeCun, & Deny, 2021) proposed a decorrelation objective that encourages the cross-correlation matrix of twin network outputs to approximate the identity, learning informative and diverse features without pairwise contrast.

With the rise of Vision Transformers, self-distillation approaches gained traction. **DINO** (Caron et al., 2021) trains a student ViT to match the outputs of a momentum teacher, producing token semantics and saliency maps “for free.” Its successor **DINO-v2** (Oquab et al., 2024) scales the recipe to 1.2 B images, yielding features competitive with supervised pre-training. In parallel,

reconstruction objectives such as **MAE** (He et al., 2022) and **SimMIM** (Xie et al., 2022) mask large portions of each image and learn to rebuild them, offering an efficient, stable alternative to contrastive learning.

2.2 Representation Learning in Medical Domain

Modern representation learning methods face unique challenges when applied to clinical data. In digital pathology, models must process gigapixel whole-slide images, often using only weak slide-level supervision and under domain shifts caused by staining variability and institutional differences (Hosseini et al., 2024). In breast ultrasound, factors like low signal-to-noise ratio, heterogeneous tissue textures, and modality-specific artifacts demand robust architectures and careful model design (Afrin et al., 2023). This section surveys how the medical imaging community has adapted learning strategies to these settings, covering multi-instance learning frameworks, pretraining approaches, and architecture comparisons tailored to the demands of pathology and ultrasound.

2.2.1 Representation Learning Challenges in Computational Pathology

Whole-slide images (WSIs) in digital pathology routinely exceed $100,000 \times 100,000$ pixels, making them several orders of magnitude larger than natural images. This extreme resolution prevents models from operating on the entire slide at once, forcing a patch-based approach that disrupts global context and complicates supervision. Moreover, annotated data is scarce, slide-level diagnostic labels are often the only signals available, and obtaining detailed annotations from pathologists is time-consuming and expensive. Finally, institutional differences in staining, scanners, and patient populations lead to significant domain shifts, limiting generalization and cross-site deployment (Hosseini et al., 2024).

A patch-wise CNN trained on the Camelyon16 benchmark demonstrated expert-level performance in lymph-node metastasis detection, establishing the viability of weakly supervised pipelines for gigapixel WSIs (Bejnordi et al., 2017). The same paradigm was later scaled to more than 44,000 WSIs from hundreds of institutions across multiple countries, reaching clinical-grade accuracy with

only slide-level labels but simultaneously exposing sharp performance drops under external validation (Campanella et al., 2019). **CLAM** introduced clustering-constrained attention into the MIL pipeline, improving data efficiency and interpretability by encouraging diverse, diagnostically relevant instance clusters (M. Y. Lu et al., 2021). **HIPT** (Hierarchical Image Pyramid Transformer) tackled the gigapixel issue architecturally by using a three-stage hierarchical ViT design that aggregates visual tokens from cell-level (16×16) to patch-level (256×256) to region-level (4096×4096), enabling multi-resolution learning while achieving computational efficiency through hierarchical feature aggregation (R. J. Chen et al., 2022). Complementing these advances, the **MIDOG 2021** challenge quantified scanner-induced domain shifts, showing that models can lose 15-25 F_1 points when applied to unseen devices, with some cases approaching 30 points, underscoring the need for stain-robust and scanner-agnostic representations (Aubreville et al., 2023).

2.2.2 Multi-Instance Learning in Computational Pathology

A single gigapixel slide can contain 10^3 – 10^6 patches, so representation learning hinges on how these instance features are aggregated into a slide prediction. **ABMIL** (Ilse, Tomczak, & Welling, 2018) introduced a learnable attention mechanism that assigns weights to each patch. To improve data efficiency and interpretability, **CLAM** (M. Y. Lu et al., 2021) augments attention with instance-level clustering, encouraging diverse yet diagnostically consistent evidence. Subsequent works enriched the aggregator itself. **DSMIL** (B. Li, Li, & Eliceiri, 2021) models instance relations to a critical highest-scored instance and couples MIL with self-supervised pre-training, while **TransMIL** (Shao et al., 2021) replaces traditional MIL attention with a Transformer encoder to capture long-range spatial correlations.

More recent methods tackle the limitations of small datasets and biased attention. **DTFD-MIL** (H. Zhang et al., 2022) distills features through a two-tier hierarchy to stabilize training on limited cohorts, and **MHIM-MIL** (Tang et al., 2023) explicitly mines hard instances, preventing the model from fixating on easy patches. Parallel work explores replacing attention with linear-time state-space layers: **S4-MIL** (Fillioux, Boyd, Vakalopoulou, Cournède, & Christodoulidis, 2023) employs a structured state-space to aggregate thousands of tokens efficiently, while **MambaMIL** (Yang, Wang, & Chen, 2024) leverages the selective-scan Mamba architecture for enhanced long sequence

modeling.

Together, these advances show a clear trend from simple attention toward richer relational modeling and more scalable sequence architectures, foundational steps in overcoming the gigapixel barrier.

2.2.3 Foundation Models for Pathology

Self-supervised pre-training has evolved from task-specific encoders to *foundation models* that generalize across diseases, organs and even modalities. An early milestone was **CTransPath** (X. Wang et al., 2022), which applied contrastive learning on millions of patches to outperform ImageNet-initialized CNNs without labels. Scaling masked ViTs further, **Phikon** (Filiot et al., 2023) used iBOT (J. Zhou et al., 2022) on 43M patches with a ViT-Base, while the public **Phikon-v2** (Filiot, Jacob, Kain, & Saillard, 2024) extractor advanced to DINOv2 (Oquab et al., 2024) with ViT-Large architecture and demonstrated strong biomarker prediction after pre-training on 460M patches. General-purpose vision models soon followed: **UNI** (R. J. Chen et al., 2024) trained on 100 M tiles from 20 tissue types and outperformed prior baselines on 34 downstream tasks, while the 632 M-parameter **Virchow** (Vorontsov et al., 2023) achieved 0.95 specimen-level AUC across common and rare cancers, with its mixed-magnification successor **Virchow2** (Zimmermann et al., 2024) advancing to state-of-the-art performance on diverse tile-level benchmarks. Complementary efforts such as **GigaPath** (Xu et al., 2024) emphasize real-world provenance, and **CHIEF** (X. Wang et al., 2024) couples diagnosis with prognosis prediction using self-supervised pre-training on ~ 15 M patches.

In parallel, vision–language models seek to align WSIs with textual domain knowledge. **Quilt-1M** (Ikezogwo et al., 2023) created the largest histopathology vision-language dataset with 1 million image-text pairs from educational YouTube videos, enabling **QUILTNET** to achieve state-of-the-art performance across multiple tasks. **Quilt-LLaVA** (Saygin Seyfioglu, Ikezogwo, Ghezloo, Krishna, & Shapiro, 2023) introduced spatially grounded instruction tuning by extracting narrators’ cursor movements, enabling reasoning beyond single image patches. **CONCH** (M. Y. Lu et al., 2024) achieved state-of-the-art performance in zero-shot classification, cross-modal retrieval, and segmentation using over 1.17 million image-caption pairs. **TITAN** (Ding et al., 2024) combines

vision-language alignment with pathology reports and synthetic captions for cross-modal retrieval and report generation. **MUSK** (Xiang et al., 2025) leverages unified masked modeling on 50 million pathology images and one billion text tokens for superior outcome prediction including melanoma relapse and immunotherapy response. **PLIP** (Z. Huang, Bianchi, Yuksekgonul, Montine, & Zou, 2023) leverages social-media pathology images to excel on rare entities, while **REMEDIS** (Azizi et al., 2023) demonstrates that combining supervised pretraining with self-supervised learning achieves data-efficient generalization.

Together, these models mark a transition from handcrafted patch pipelines to versatile, pre-trained backbones, whether vision-only or multimodal, that can be fine-tuned or prompted for diverse clinical tasks, setting the stage for unified representation learning in computational pathology.

2.2.4 Deep Learning in Breast Ultrasound Imaging

Two public benchmarks have driven progress in ultrasound breast cancer analysis. "Dataset B" provides lesion annotations for automated detection and classification (Yap et al., 2017), whereas BUSI offers a three-class task that separates normal, benign and malignant images (Al-Dhabyani, Gomaa, Khaled, & Fahmy, 2020). Early CNN work transferred ImageNet features to ultrasound; a color-converted VGG19 (Simonyan & Zisserman, 2015) achieved an 84% accuracy on Dataset B and confirmed that generic visual cues remain informative despite speckle noise (Byra et al., 2019). Weak-label settings followed when a weakly-supervised approach matched fully supervised baselines using only image-level malignancy labels (Kim et al., 2021).

Transformer encoders soon entered the field. A Vision Transformer fine-tuned on BUSI reached parity with ResNet-50 (Gheflati & Rivaz, 2022). Hierarchical self-attention improved robustness further when a Swin-based variant incorporated multi-scale context (C. Zhu et al., 2024). Ensemble methods then combined heterogeneous CNNs in a meta-learner that improved the F_1 score on BUSI (Ali et al., 2023).

These advances show clear progress from basic transfer learning to sophisticated multi-scale architectures and ensemble methods.

2.3 Representation Engineering and Interpretability

As deep learning evolved from convolutional networks to transformers and foundation models, a critical challenge emerged: understanding what these models learn and how they decide. This is particularly important in medical imaging, where clinical adoption requires accurate predictions and interpretable explanations that clinicians can validate and trust.

The field has responded with two approaches: post-hoc interpretability methods that explain trained models through visualization and attribution techniques, and interpretable-by-design approaches that build transparency into architectures through sparse representations and mechanistic constraints. This section examines both directions, from classical gradient-based methods to modern sparse autoencoders, highlighting their capabilities and limitations in creating models that are both powerful and interpretable.

2.3.1 Classical Computer-Vision Interpretability Methods and Their Limitations

Early computer-vision interpretability focused on **gradient-based attribution**, which back-propagates a network’s output to the input pixels. **Saliency Map** (Simonyan et al., 2013) computed the gradient of the class score with respect to the input image, producing coarse visual explanations. Subsequent variants improved stability or theoretical grounding. **Guided Backpropagation** (Springenberg, Dosovitskiy, Brox, & Riedmiller, 2014) filtered negative gradients to produce sharper maps, while **SmoothGrad** (Smilkov, Thorat, Kim, Viégas, & Wattenberg, 2017) reduced noise by averaging attributions over multiple noisy inputs. **Integrated Gradients** (Sundararajan, Taly, & Yan, 2017) introduced path-based integration to address saturation and baseline dependence. **DeepLIFT** (Shrikumar, Greenside, & Kundaje, 2017) and **Layer-wise Relevance Propagation** (Bach et al., 2015) redistributed prediction scores through network layers with conservation constraints. **Deep Taylor Decomposition** (Montavon, Lapuschkin, Binder, Samek, & Müller, 2017) further formalized these ideas by decomposing outputs into input contributions via Taylor expansions.

Complementary to derivative views, **class-activation-map** methods localize discriminative regions by combining spatial and channel statistics. **CAM** (B. Zhou, Khosla, Lapedriza, Oliva, &

Torralla, 2016) linearly combined convolutional feature maps weighted by the final classifier, but required architectural constraints like global average pooling. **Grad-CAM** (Selvaraju et al., 2017) generalized the approach using gradient weights, enabling broader applicability. Successors such as **Grad-CAM++** (Chattopadhyay, Sarkar, Howlader, & Balasubramanian, 2018) improved multi-object handling, while **Score-CAM** (H. Wang et al., 2020) eliminated reliance on gradients altogether. **XGrad-CAM** (Fu et al., 2020) introduced axiomatic constraints, and **Ablation-CAM** (Ramaswamy et al., 2020) estimated feature importance via occlusion. **Eigen-CAM** (Muhammad & Yeasin, 2020) applied PCA to activations, and **LayerCAM** (Jiang, Zhang, Hou, Cheng, & Wei, 2021) extended attribution to intermediate layers. (Chefer et al., 2021b) adapted these principles to attention-based architectures such as ViT (Dosovitskiy et al., 2021).

A parallel family substitutes gradients with **perturbation-based surrogates**, estimating attribution by measuring output changes under input modifications. (Zeiler & Fergus, 2014) slid a masking window across the image to observe class score drops. **LIME** (Ribeiro, Singh, & Guestrin, 2016) locally approximated the model with sparse linear regressions on perturbed samples. **SHAP** (Lundberg & Lee, 2017) combined Shapley values with model-agnostic approximations for principled feature attributions. **RISE** (Petsiuk, Das, & Saenko, 2018) used random masks and weighted model responses to estimate importance maps. More recent variants such as **Anchors** (Ribeiro, Singh, & Guestrin, 2018) returned high-precision rule sets, and **L2X** (J. Chen, Song, Wainwright, & Jordan, 2018) learned to select informative input subsets in a single forward pass via mutual information maximization.

2.3.2 Sparse Representations

Early theories of efficient coding proposed that natural images can be represented by a small number of active basis elements, a principle formalized as **Sparse Coding** (Olshausen & Field, 1997). This insight gave rise to overcomplete dictionary learning methods such as **K-SVD** (Aharon, Elad, & Bruckstein, 2006), which generalizes K-means through alternating sparse coding and SVD-based dictionary updates; **Efficient Sparse Coding Algorithms** (Lee, Battle, Raina, & Ng, 2006), which introduced the feature-sign search and Lagrange dual methods to accelerate inference; and **Online Dictionary Learning** (Mairal, Bach, Ponce, & Sapiro, 2009), which leveraged stochastic

approximations to scale learning to large datasets. These approaches share a common goal: to learn compact, sparse representations that preserve essential structure while promoting **interpretability** through selectivity and efficiency.

Building on these principles, deep networks began to incorporate sparsity as an architectural prior. In particular, **Deep Sparse Rectifier Networks** (Glorot, Bordes, & Bengio, 2011) demonstrated that ReLU activations naturally yield sparse hidden activations. Expanding on this idea, **k -Sparse Autoencoders** (Makhzani & Frey, 2013) introduced an explicit top- k constraint on hidden units to enforce sparsity directly. These **Sparse Autoencoders** (SAEs) aim to represent inputs through a distributed but selective code, where only a small number of neurons are active for any given input, thereby aligning with **interpretability** objectives. The resulting representations tend to localize features and disentangle factors of variation, making individual units easier to inspect and understand than dense, entangled embeddings.

In parallel to sparsity-based approaches, another line of work pursued the goal of **disentangled representations**, where each latent factor captures a distinct, interpretable concept. Learning disentangled structure became a parallel goal. β -VAE (Higgins et al., 2017) encouraged factorized latent variables via a scaled KL penalty, **InfoGAN** (X. Chen et al., 2016) maximized mutual information between latent codes and generated samples, and **Network Dissection** (Bau, Zhou, Khosla, Oliva, & Torralba, 2017) quantified unit-level interpretability by aligning activations with human-labeled concepts. These frameworks treat sparsity and disentanglement as complementary routes to semantic feature discovery.

Recent work increasingly treats **sparsity** as a deliberate engineering tool for mechanistic transparency. **Toy Models of Superposition** (Elhage et al., 2022) showed that dense neurons multiplex features, motivating dictionaries that isolate single-concept directions. **Monosemantic Sparse Autoencoders** (Bricken et al., 2023) decomposed Transformer activations into near one-to-one feature–concept pairs. **Gated SAEs** (Rajamanoharan et al., 2024) address a core limitation of traditional SAEs—shrinkage bias introduced by the ℓ_1 penalty—by decoupling feature selection from magnitude estimation and applying sparsity only to the gating layer. This achieves similar interpretability with half as many active features and improved reconstruction fidelity. **BatchTopK SAEs** (Bussmann, Leask, & Nanda, 2024) replace the sample-level Top- k constraint with a batch-level selection

mechanism, selecting the top $n \times k$ activations across a batch of n samples to allow more adaptive latent allocation while preserving average sparsity. Finally, (Gao et al., 2025) trains SAEs with up to 16 million latents on GPT-4 activations, reveals smooth scaling laws, introduces new interpretability metrics, and eliminates dead latents through architectural refinements. Together, these advances frame sparsity not merely as a regularizer but as a scalable, practical mechanism for building interpretable, monosemantic internal representations.

2.4 Aligned Representations

While interpretability methods enable understanding of individual models, the rapid development of diverse architectures has created a scalability problem. Unlike standard metrics where we can benchmark dozens of models by running each once to obtain comparable scores, interpretability analysis requires analyzing each model individually, extracting complex insights about learned representations, and manually comparing findings across architectures. This process becomes exponentially more expensive as the number of models grows, creating a critical bottleneck that prevents understanding concept representation at the pace of model development. Furthermore, we discuss some vision-language models as an examples of aligned representations that bridge different modalities in shared embedding spaces.

2.4.1 Cross-Model Representation Engineering

While sparse representations enable interpretability within a single model, the growing need to compare models across architectures and training regimes demands cross-model tools to align, audit, and interpret representations in a unified space.

Initial approaches to this problem focused on measuring **representational similarity**. **Convergent Learning** revealed that even when trained identically, different networks learn only partially overlapping feature spaces, highlighting the need for formal tools to assess alignment (Y. Li, Yosinski, Clune, Lipson, & Hopcroft, 2016). **SVCCA** addressed this by projecting activations into low-dimensional canonical subspaces invariant to affine transformations, enabling direct comparison across layers and models (Raghu, Gilmer, Yosinski, & Sohl-Dickstein, 2017). **PWCCA** refined

this further by weighting directions according to their contribution, filtering out noisy or unstable components (Morcos, Raghu, & Bengio, 2018). Later, **CKA** introduced a similarity index with improved invariance properties and interpretability, becoming a standard metric for cross-architecture comparison (Kornblith, Norouzi, Lee, & Hinton, 2019). These tools revealed, for example, that Vision Transformers develop more homogeneous and globally consistent representations than CNNs, which exhibit layer-wise progression and localized features (Raghu, Unterthiner, Kornblith, Zhang, & Dosovitskiy, 2021).

Beyond measuring similarity, more recent methods aim to **functionally align** networks. **Model Stitching** connects the lower layers of one model to the upper layers of another via a trainable adapter, testing whether intermediate representations are interchangeable (Bansal, Nakkiran, & Barak, 2021). Follow-up work extended this to heterogeneous architectures by using techniques such as strided convolutions and upsampling, allowing comparison across differing widths and depths (Hernandez, Dangovski, & Lu, 2022). These studies show that even structurally distinct models may converge toward compatible latent features under shared training objectives.

A more recent line of work reframes the problem as one of **representation engineering**. Rather than comparing models pairwise or relying on adapters, **Universal Sparse Autoencoders (USAE)** construct a shared, overcomplete dictionary trained on activations from multiple models (Thasarathan, Forsyth, Fel, Kowal, & Derpanis, 2025). Each model is encoded into sparse, interpretable vectors within this common space, enabling direct comparison and concept alignment across architectures. By enforcing monosemanticity and sparsity, USAE supports interpretable inspection and editing without re-analyzing each model in isolation.

2.4.2 Multimodal Vision-Language Alignment

Aligned representations are a core principle in vision-language modeling, where the goal is to embed images and text into a shared semantic space. This joint embedding allows simple similarity measures to support a wide range of downstream tasks, including retrieval, zero-shot classification, captioning, and multimodal reasoning. Unlike model-model alignment, which compares internal activations across architectures, multimodal alignment bridges fundamentally different input types. The result is a unified space where distance reflects semantic similarity regardless of modality, a

key enabler of flexible, interpretable AI systems.

CLIP (Radford et al., 2021) established this paradigm using a dual-encoder architecture and a contrastive loss over 400 million image–text pairs, achieving strong zero-shot performance across diverse tasks. **ALIGN** (Jia et al., 2021) scaled this approach with 1.8 billion noisy pairs, improving robustness and retrieval accuracy. Prior fusion-based models such as **ViLBERT** (J. Lu, Batra, Parikh, & Lee, 2019), **LXMERT** (H. Tan & Bansal, 2019), and **UNITER** (Y.-C. Chen et al., 2020) used cross-attention within shared transformers to model fine-grained relationships between tokens and regions, enabling strong performance on VQA and captioning. More recently, **SimVLM** (Z. Wang et al., 2021) showed that a unified transformer trained with a generative prefix language modeling objective can also learn effective joint representations, without relying on contrastive losses. Across these architectures, the shared embedding space remains central, serving as the foundation for modality-agnostic transfer and interpretable alignment across vision and language.

2.5 Concluding Remarks

This review covers visual representation learning from early CNNs to modern foundation models, showing consistent progress toward more efficient and scalable architectures. The evolution from local convolutions to global attention to selective state updates shows how each generation addresses its predecessor’s limitations while introducing new capabilities. Self-supervised learning has transformed the field by enabling models to use vast unlabeled datasets, while foundation models show that large-scale pretraining yields broadly transferable representations.

Yet when we try to apply these advances to real-world problems, significant limitations emerge. Medical imaging exposes the clearest gaps: computational pathology struggles with gigapixel resolutions and weak supervision, while ultrasound analysis faces low signal-to-noise ratios and heterogeneous tissue textures. Both domains also suffer from domain shifts across institutions and equipment. These challenges become critical when clinical adoption requires both accuracy and interpretability, something our current methods don’t deliver.

The interpretability problem extends beyond medical domains. Gradient-based and perturbation

techniques provide simple explanations, but we remain uncertain whether they represent what models actually learn. Meanwhile, the growth of diverse architectures creates a scalability bottleneck: we can't analyze each new model individually.

Cross-model alignment offers a potential solution by enabling comparison across architectures in shared spaces. However, current alignment approaches remain limited in scope and scalability, unable to handle the range of architectures and modalities we're building.

These three gaps, efficient medical imaging representations, scalable interpretability, and robust alignment methods, drive the research contributions in the following chapters.

Chapter 3

Efficient Representation Learning in Medical Domain

3.1 Introduction: Why Efficient Representation Learning?

The deployment of deep learning models in clinical settings faces a fundamental trade-off between computational demands and practical constraints. While modern architectures like Vision Transformers achieve impressive performance on natural image benchmarks, their quadratic complexity and substantial memory requirements create significant barriers to adoption in medical domains where computational resources are often limited and real-time processing is essential for clinical workflows.

Medical imaging presents unique computational challenges that worsen these efficiency concerns. In computational pathology, whole-slide images routinely exceed $100,000 \times 100,000$ pixels, requiring models to process thousands of patches per slide while maintaining global coherence under weak supervision constraints. In ultrasound imaging, models must operate on resource-constrained devices while handling low signal-to-noise ratios and heterogeneous tissue textures. These domain-specific requirements demand architectures that achieve strong performance without prohibitive computational overhead.

Traditional approaches to this efficiency challenge have involved either accepting reduced model

capacity (using smaller CNNs (LeCun et al., 2002)) or implementing complex hierarchical processing pipelines. However, CNNs fundamentally struggle with long-range dependencies crucial for understanding tissue organization and cellular neighborhoods, while hierarchical approaches introduce additional complexity and potential information loss. Vision Transformers (Dosovitskiy et al., 2021), despite their superior modeling capacity, remain computationally prohibitive for many clinical applications due to their $\mathcal{O}(L^2)$ attention complexity.

The emergence of state-space models, particularly the Mamba architecture, offers a compelling alternative that addresses both efficiency and performance requirements. By combining linear computational complexity with the ability to model long-range dependencies, Vision Mamba architectures promise to bridge the gap between computational constraints and clinical needs. The selective state-space mechanism enables input-dependent processing that can adapt to the varying information density characteristic of medical images, while maintaining the sequential inductive biases that prove valuable for modeling spatial relationships in pathology tissues.

This chapter investigates how Vision Mamba architectures perform in medical representation learning through two domains of computational pathology and ultrasound imaging.

Section 3.2 presents Vim4Path (Nasiri-Sarvi, Trinh, Rivaz, & Hosseini, 2024), our work on self-supervised Vision Mamba (L. Zhu et al., 2024) for histopathology images, where we show how DINO-based training (Caron et al., 2021) of Vision Mamba models achieves superior performance compared to Vision Transformers, particularly in parameter-constrained settings. This case study reveals how sequential processing naturally aligns with pathologist workflows and enables more effective patch-to-slide aggregation under weak supervision constraints.

Section 3.3 examines Vision Mamba for classification of breast ultrasound images (Nasiri-Sarvi, Hosseini, & Rivaz, 2024), demonstrating how VMamba (Y. Liu et al., 2024) architectures can effectively transfer from natural image domains to medical imaging tasks. Through comprehensive evaluation across multiple ultrasound datasets and rigorous statistical analysis, we show that Mamba-based models achieve statistically significant performance improvements.

Our findings reveal that efficient representation learning in medical domains requires more than computational optimization. It demands architectures that align with domain-specific processing

patterns. The sequential scanning patterns of Vision Mamba naturally mirror pathologist workflows in tissue examination, while the selective attention mechanisms prove particularly effective for highlighting diagnostically relevant features in noisy ultrasound images.

The implications extend beyond immediate performance gains to broader questions of model interpretability and clinical integration. As we demonstrate through explainability analyses, efficient architectures can simultaneously improve both computational performance and alignment with expert knowledge, creating a foundation for trustworthy clinical decision support systems. Section 3.4 synthesizes these findings and discusses how the alignment between efficiency and interpretability motivates the subsequent exploration of representation engineering and cross-model analysis presented in later chapters.

3.2 Self-Supervised Vision Mamba for Histopathology Images

Computational pathology faces fundamental challenges when processing gigapixel whole-slide images (WSIs). Each WSI comprises hundreds of thousands of cells with complex biological interactions, requiring extraction of thousands of manageable patches for analysis under weak supervision constraints. Traditional approaches using ImageNet-pretrained features suffer from domain distribution mismatch, while Vision Transformers, despite their modeling capacity, require substantial computational resources that limit clinical deployment. This section presents Vim4Path (Nasiri-Sarvi, Trinh, et al., 2024), our approach leveraging Vision Mamba architectures within the DINO self-supervised learning framework to address these efficiency and performance challenges.

3.2.1 Methodology

Vision Mamba Architecture for Pathology. Vision Mamba (L. Zhu et al., 2024) combines the localized inductive biases of CNNs with the long-range dependency modeling capabilities of Vision Transformers while maintaining linear computational complexity. As illustrated in Figure 3.1, unlike CNNs that are constrained by kernel size for capturing dependencies between neighboring segments (A and B), and ViTs that treat all patches equally regardless of spatial proximity, Vision

Mamba employs sequential processing that naturally captures both short-range and long-range relationships (from segment A to segment Z) through its selective state-space mechanism based on Mamba (Gu & Dao, 2024).

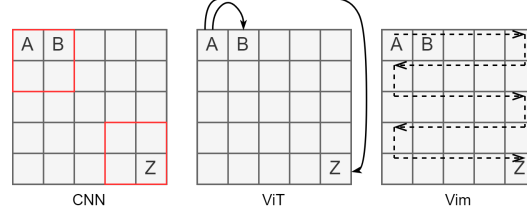


Figure 3.1: Comparison between different architecture designs. Vim sequential processing allows the model to capture both short-range and long-range dependencies.

The detailed architecture within the DINO framework is shown in Figure 3.2. We adapted the standard Vision Mamba architecture by incorporating positional encoding interpolation to handle varying image sizes required by the DINO framework, allowing the model to process both global views (224×224) and local views (96×96) effectively. The architecture processes image patches through a bidirectional scanning pattern, enabling the model to capture dependencies between neighboring tokens while maintaining global context awareness.

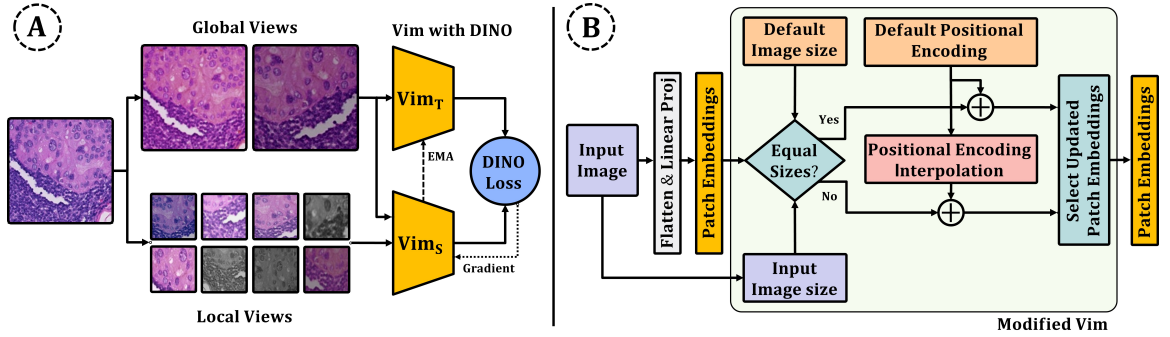


Figure 3.2: Detailed architecture of VIM within the DINO framework. We modify the Vim model to adapt to input image size for positional embedding interpolation and employ the modified model within DINO as a backbone architecture for self-supervised learning.

DINO-based Self-Supervised Learning. We employ DINO (Caron et al., 2021) (Self-Distillation

with No Labels) as our self-supervised learning framework, which utilizes a teacher-student setup where both networks share identical architecture but differ in parameters. The framework crops input images into global and local views with various augmentations. The student network predicts teacher outputs from all views while the teacher processes only global views. Teacher parameters are updated through exponential moving average of student weights, ensuring training stability.

This global-local view training scheme proves particularly valuable for pathology images, where understanding transitions between healthy and cancerous states requires modeling relationships across multiple scales, from individual cells to tissue neighborhoods. The DINO framework’s emphasis on multi-scale relationships aligns naturally with pathologist workflows that examine tissues from local cellular details to broader architectural patterns.

Pathological Relevance of Sequential Processing. The sequential scanning pattern of Vision Mamba uniquely aligns with clinical pathology practices, as depicted in Figure 3.3. During cancer evolution, normal cells undergo coordinated responses with neighboring cells, creating sequential patterns that interact with mutation-burdened tumor cells (Quail & Joyce, 2013). Each patch is divided into multiple tokens, and Vision Mamba raster scans the tokens following a lawnmower pattern, enabling the encoder to capture short-range dependencies for representation in patch embedding. This process imitates the realistic representation of cancer cells in sequence, from local zooming to global representation on the slide, where similar slide navigation is used by pathologists for diagnosing cancer cells under microscopy (Molin et al., 2015).

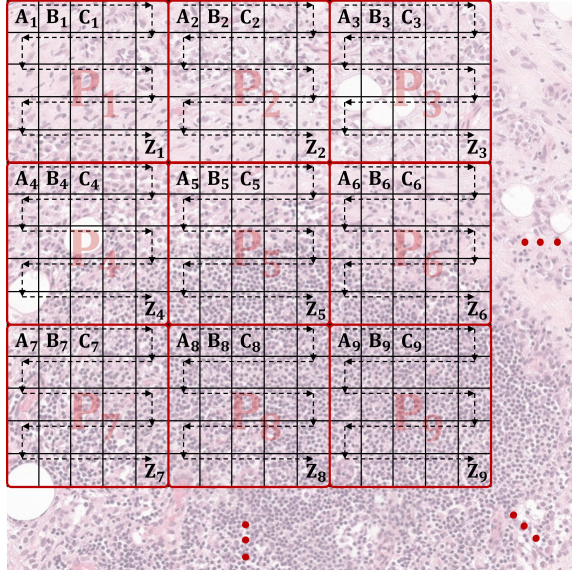


Figure 3.3: Sequential processing of Vim done on each patch level from slide for feature embedding. This is similar to the lawnmower pattern used for slide navigation by pathologists to study cellular neighbourhoods in the tissue for cancer diagnosis. The information from each patch (i.e. embeddings) are put together to reach to a consensus on the slide level (i.e. aggregation).

3.2.2 Experiments

Dataset and Preprocessing. We evaluated our approach on the Camelyon16 dataset (Bejnordi et al., 2017), which provides whole-slide images of lymph node sections for metastasis detection. As shown in Table 3.1, the training set contains 158 normal and 110 tumor slides, while the test set includes 80 normal and 49 tumor slides.

Table 3.1: Camelyon16 dataset distribution

Category	Training	Testing
Normal slides	158	80
Tumour slides	110	49

For self-supervised pretraining, we extracted patches at 5× and 10× magnification levels using the CLAM preprocessing pipeline (M. Y. Lu et al., 2021). Table 3.2 summarizes the distribution of

extracted patches, yielding 167,954 patches at 5× (146,508 normal, 21,446 tumor) and 1,001,839 patches at 10× magnification (555,231 normal, 446,608 tumor). Note that tumor patches are extracted from tumor slides without reference to region-of-interest labels, indicating not all extracted patches necessarily represent tumor regions.

Table 3.2: Pre-training patch distribution (tumor patches come from tumor slides without ROI filtering)

Zooming level	Normal patches	Tumor patches
5×	146,508	21,446
10×	555,231	446,608

To enable rigorous patch-level evaluation, we created balanced datasets (PCam-224-5× and PCam-224-10×) by extracting patches from tumor regions with at least 50% intersection with region-of-interest annotations, supplemented with uniformly sampled normal patches. Table 3.3 shows the resulting balanced datasets with expert pathologist validation ensuring label accuracy.

Table 3.3: PCam patch-classification dataset (balanced classes: counts shown as $2 \times N$)

Dataset	Training patches	Testing patches
PCam-224-5×	$2 \times 7\,602$	$2 \times 6\,611$
PCam-224-10×	$2 \times 28\,955$	$2 \times 25\,370$

Model Configurations and Training. We compared Vision Mamba models against Vision Transformer baselines across multiple scales, as detailed in Table 3.4: ViT-tiny (6M parameters) versus Vim-tiny (7M parameters), and ViT-small (22M parameters) versus Vim-small (26M parameters). Additionally, we evaluated Vim-tiny-plus (13M parameters) with increased embedding dimension to investigate architectural trade-offs.

Table 3.4: Trainable parameters (rounded to the nearest million)

Model	Parameters (M)
ViT-ti	6
ViT-s	22
Vim-ti (ours)	7
Vim-ti-plus (ours)	13
Vim-s (ours)	26

Training utilized 4 NVIDIA V100 GPUs with an effective batch size of 512, training each model for 100 epochs within the DINO framework. We employed cosine learning rate scheduling with linear warmup, starting from learning rate 0.0005 with weight decay increasing from 0.04 to 0.4 using cosine scheduling. For slide-level classification, we used the CLAM-SB pipeline (M. Y. Lu et al., 2021) as the multi-instance learning aggregator.

3.2.3 Results

Slide-Level Classification Performance. Vision Mamba models demonstrated substantial improvements over Vision Transformer counterparts, particularly at smaller scales. Table 3.5 presents comprehensive results showing that Vim-tiny achieved ROC AUC of 95.81 compared to ViT-tiny’s 87.60, an improvement of 8.21 AUC points with comparable parameter count. This enhancement underscores Vim’s efficacy in resource-constrained settings.

When considering scaled models, Vim-tiny-plus (13M parameters) outperformed the larger ViT-small (22M parameters) by achieving AUC of 97.39 versus 96.76, demonstrating efficiency advantages. Furthermore, scaling to Vim-small (26M parameters) achieved the highest AUC of 98.85 among all models, emphasizing consistent performance advantages across scales.

Patch-Level Classification Performance. For patch-level evaluation, we assessed models on PCam-224 datasets using linear evaluation protocol. Table 3.6 shows results for both PCam-224-10× and PCam-224-5×. On the 10× dataset, Vim-tiny achieved 95.99% accuracy compared to ViT-tiny’s 94.59%. Similarly, on the 5× dataset, Vim-tiny achieved 90.45% accuracy versus ViT-tiny’s 88.87%.

Table 3.5: Slide-level classification on Camelyon16. AUC is the primary comparison metric.

Aggregator model	Zoom	Encoder	Pre-train set	Method	Params (M)	ACC	F1	AUC
Max-Pooling	20×	ResNet-50	ImageNet	Sup.	25	78.95	71.06	81.28
Mean-Pooling	20×	ResNet-50	ImageNet	Sup.	25	76.69	70.41	80.07
AB-MIL (Ilse et al., 2018)	20×	ResNet-50	ImageNet	Sup.	25	90.06	87.40	94.00
DSMIL B. Li et al. (2021)	20×	ResNet-50	ImageNet	Sup.	25	90.17	87.65	94.57
CLAM-SB M. Y. Lu et al. (2021)	20×	ResNet-50	ImageNet	Sup.	25	90.31	87.89	94.65
CLAM-MB M. Y. Lu et al. (2021)	20×	ResNet-50	ImageNet	Sup.	25	90.14	88.10	94.70
TransMIL Shao et al. (2021)	20×	ResNet-50	ImageNet	Sup.	25	89.22	85.10	93.51
DTFD-MIL H. Zhang et al. (2022)	20×	ResNet-50	ImageNet	Sup.	25	90.22	87.62	95.15
MHIM-MIL Tang et al. (2023)	20×	ResNet-50	ImageNet	Sup.	25	92.48	90.75	96.49
HIPT R. J. Chen et al. (2022)	20×	ViT-s	TCGA	DINO	22	NA	NA	95.7
iBot-COAD Filiot et al. (2023)	20×	ViT-B	TCGA-COAD	iBot	86	NA	NA	94.5
CTransPath X. Wang et al. (2022)	20×	CNN	TCGA + PAIP	Swin	NA	92.2	NA	94.2
CLAM-SB M. Y. Lu et al. (2021)	10×	ViT-ti	Cam16	DINO	6	90.69	86.36	87.60
CLAM-SB M. Y. Lu et al. (2021)	10×	Vim-ti (ours)	Cam16	DINO	7	93.02	90.32	95.81
CLAM-SB M. Y. Lu et al. (2021)	10×	Vim-ti+ (ours)	Cam16	DINO	13	93.79	91.83	97.39
CLAM-SB M. Y. Lu et al. (2021)	10×	ViT-s	Cam16	DINO	22	94.57	92.47	96.76
CLAM-SB M. Y. Lu et al. (2021)	10×	Vim-s (ours)	Cam16	DINO	26	92.24	89.79	98.85

Table 3.6: Linear-evaluation accuracy on PCam-224 datasets.

Dataset	ViT-ti	Vim-ti (ours)	Vim-ti+ (ours)	ViT-s
PCam-10×	94.59	95.99	96.48	96.65
PCam-5×	88.87	90.45	90.39	90.36

Notably, Vim-tiny-plus consistently outperformed ViT-small with approximately half the parameters across both magnification levels, demonstrating the architecture’s parameter efficiency for pathology applications.

Ablation Studies. Table 3.7 demonstrates the impact of magnification level on performance, showing that models pretrained at 10× magnification achieve superior results compared to those trained at 5×, primarily due to the larger pretraining dataset available at higher magnification.

Table 3.7: Effect of zooming level on MIL.

Pre-trained weights	Model	Cam16 MIL 5×		Cam16 MIL 10×	
		AUC	ACC	AUC	ACC
Cam16-5× DINO	ViT-ti	65.91	73.64	68.85	65.11
	Vim-ti (ours)	71.30	73.64	64.41	69.76
	Vim-ti-plus (ours)	79.03	75.96	74.08	72.86
	ViT-s	67.72	76.74	69.87	74.41
Cam16-10× DINO	ViT-ti	75.05	72.86	87.60	90.69
	Vim-ti (ours)	62.14	54.26	95.81	93.02
	Vim-ti-plus (ours)	71.47	79.06	97.39	93.79
	ViT-s	84.89	81.39	96.76	94.57
	Vim-s (ours)	79.64	79.06	98.85	92.24

Table 3.8 reveals that feature representations learned at one magnification can transfer effectively to other levels in patch classification tasks, with models pretrained on larger 10× datasets demonstrating superior representation quality during linear evaluation.

Table 3.8: Effect of zooming level on patch classification (linear evaluation).

Pre-trained weights	Model	5× ACC	10× ACC
Cam16-5× DINO	ViT-ti	88.87	89.29
	Vim-ti (ours)	90.45	91.03
	Vim-ti-plus (ours)	90.39	88.72
	ViT-s	90.36	89.92
Cam16-10× DINO	ViT-ti	90.99	94.59
	Vim-ti (ours)	93.01	95.99
	Vim-ti-plus (ours)	93.23	96.48
	ViT-s	93.00	96.65
	Vim-s (ours)	93.04	96.51

3.2.4 Explainability Analysis

We conducted explainability analysis using GradCAM (Selvaraju et al., 2017) to generate activation heatmaps from CLS tokens, despite both models being trained without labels. Expert pathologist review revealed distinct attention patterns between architectures.

Figure 3.4 shows representative Vim-small heatmaps that consistently focused on cancer-specific cellular features and interfaces with non-cancer cells. Vim highlighted intracellular mucin (red asterisks), a feature nearly 100% specific for cancer, and adjacent activated lymphocytes (yellow asterisks) that are biologically reactive to cancer presence. Both features are exclusively found in cancer contexts.

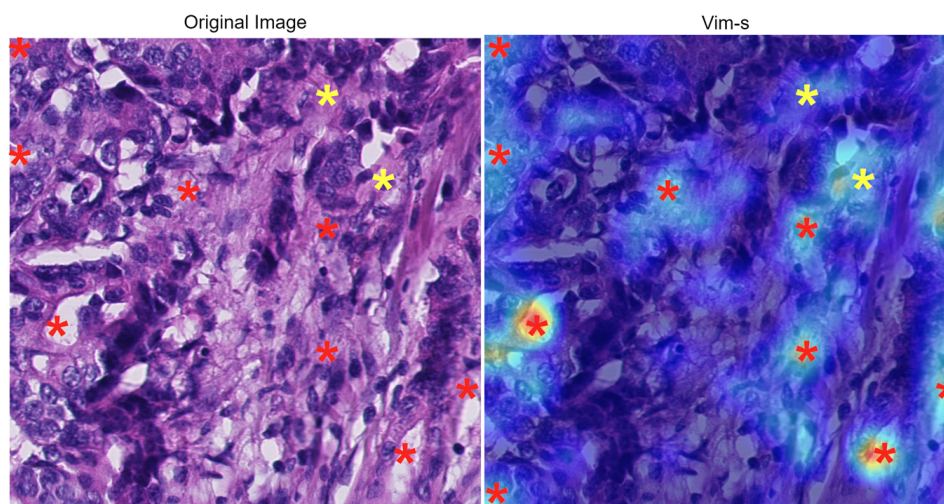


Figure 3.4: Representative tumor patch with Vim-s heatmap. The red asterisks highlight intracellular mucin in cancer cells. The yellow asterisks highlight stromal features adjacent to cancer cells. (The heatmaps are generated at 10x and overlaid on 40x images.)

In contrast, Figure 3.5 demonstrates ViT-small heatmaps that focused primarily on cancer cells as whole entities, particularly atypical cancer cells, in a more dichotomous cancer versus non-cancer fashion (red asterisks). While both models performed effectively, Vim’s ability to identify specific biological features used by pathologists for cancer validation suggests closer alignment with expert diagnostic workflows, indicating that the proposed Vim4Path framework mimics the workflow of pathologists.

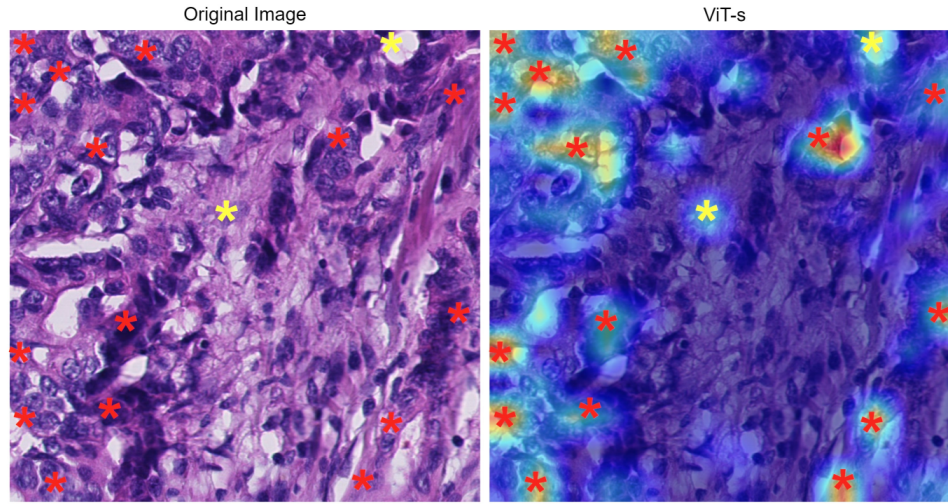


Figure 3.5: Representative tumor patch with ViT-s heatmap. The red asterisks highlight areas centralized on cancer cells. The yellow asterisks highlight other features, notably a focus of intracellular mucin (top-right) and a stromal cell (middle). (The heatmaps are generated at 10x and overlaid on 40x images.)

3.2.5 Discussion

Our findings reveal that Vision Mamba architectures offer compelling advantages for computational pathology applications. The substantial performance improvements at smaller scales (8.21 AUC point gain for comparable parameter counts) demonstrate practical benefits for resource-constrained clinical environments. The architecture’s sequential processing naturally aligns with pathologist workflows, as evidenced by explainability analysis showing focus on diagnostically relevant cellular features rather than whole-cell dichotomous patterns.

The performance advantages extend beyond computational metrics to interpretability alignment with domain expertise. Vim’s attention to intracellular mucin and tumor-adjacent lymphocytes reflects the sequential cellular neighborhood analysis that pathologists employ for diagnostic validation. This alignment between computational efficiency and clinical workflow integration suggests promising directions for practical deployment in diagnostic settings, particularly given the medical

and biological need for computational pathology algorithms that query neighboring cellular dependencies.

However, our study faced computational constraints that limited extensive hyperparameter optimization and longer training regimes. The compute constraints prevented exhaustive hyper-parameter tuning and more extended training regimes, which could unlock even higher performance from the Vim model. Future work should explore scaling to 20× magnification, the clinical standard, and evaluation across diverse pathology datasets to validate generalization beyond lymph node metastasis detection.

3.3 Vision Mamba for Classification of Breast Ultrasound Images

This section presents our investigation of Vision Mamba architectures for breast ultrasound classification (Nasiri-Sarvi, Hosseini, & Rivaz, 2024), demonstrating how VMamba and Vim models can effectively transfer from natural image domains to medical ultrasound applications while maintaining the computational efficiency crucial for clinical deployment.

3.3.1 Methodology

Adaptation of Vision Mamba Architectures. We employed two primary Vision Mamba variants adapted for ultrasound classification: Vim (L. Zhu et al., 2024) and VMamba (Y. Liu et al., 2024). As illustrated in Figure 3.6, Vim processes images by dividing them into smaller patches, each projected into patch embeddings. A bidirectional Mamba processes these patches by considering both previous and next tokens, with positional encoding added to retain spatial information about neighboring patches. This approach mirrors ViT processing but substitutes attention-based Transformer blocks with Mamba-based blocks, enabling linear computational complexity while maintaining global context awareness.

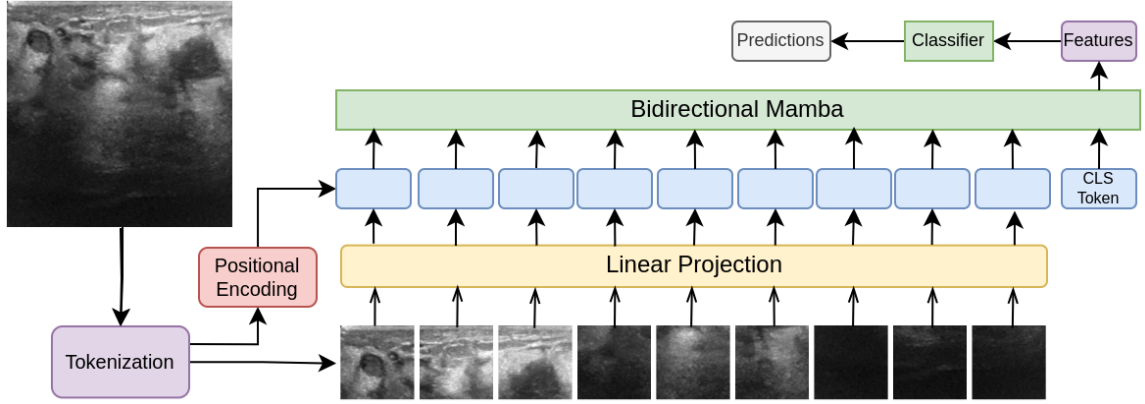


Figure 3.6: Overview of the Vim model architecture for ultrasound image processing.

VMamba introduces a more sophisticated approach through "2D Selective Scan," as shown in Figure 3.7. Instead of breaking images into tokens like ViT and Vim, VMamba treats image patches as feature maps processed using VSS blocks similar to CNN models, where feature maps are down-sampled at each layer. This hierarchical processing enables multi-scale feature extraction while maintaining the selective state-space benefits of the Mamba architecture.

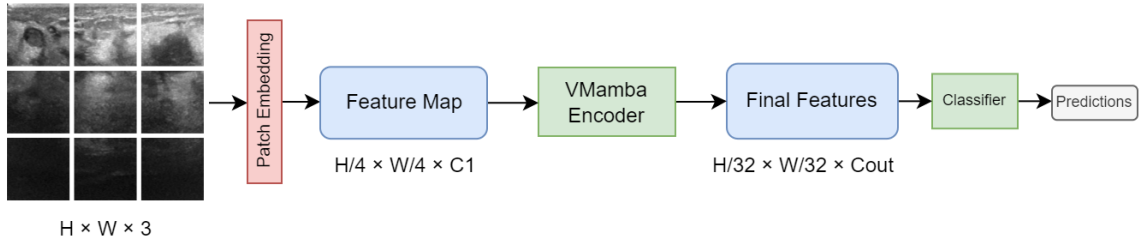


Figure 3.7: Overview of the VMamba model architecture with 2D Selective Scan mechanism.

Architectural Advantages for Ultrasound Processing. Figure 3.8 illustrates the fundamental differences between CNN, ViT, and Mamba-based processing approaches. CNNs struggle with long-range dependencies when patches P1 and P2 fall outside the same receptive field. ViTs can process relationships between distant patches (P1-P9) through attention mechanisms, but distinguishing between close-range (P1-P2) and long-range (P1-P9) dependencies relies primarily on positional encoding, requiring substantial training data to learn these distinctions effectively.

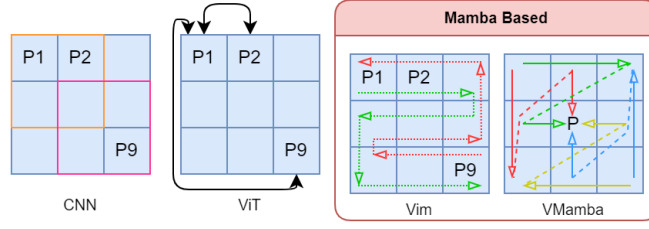


Figure 3.8: Abstract comparison between different architecture types for processing spatial relationships in ultrasound images.

Mamba-based models reintroduce inductive bias through sequential processing while maintaining long-range information processing capabilities. This combination proves particularly valuable for ultrasound imaging, where understanding both local tissue characteristics and global anatomical context is essential for accurate classification. Vim employs bidirectional scanning for multi-directional feature extraction, while VMamba’s 2D selective scan captures more comprehensive relationships through four-directional processing, enhancing feature representation and contextual learning.

Transfer Learning Strategy. Given the limited size of ultrasound datasets, we employed transfer learning from ImageNet-pretrained weights. This approach leverages the rich feature representations learned on natural images and adapts them to the specific characteristics of ultrasound imaging through fine-tuning. The selective state-space mechanism of Mamba architectures proves particularly effective for this transfer, as the input-dependent processing can adapt to the distinct statistical properties of ultrasound images while maintaining the structural knowledge from natural image pretraining.

3.3.2 Experimental Setup

Datasets and Preprocessing. We evaluated our approach on three configurations: the BUSI dataset (Al-Dhabyani et al., 2020) for three-class classification (normal, benign, malignant), the B dataset (Yap et al., 2017) for two-class classification (benign, malignant), and a combined BUSI+B dataset for comprehensive three-class evaluation. Each experiment utilized transfer learning with

ImageNet-pretrained weights to address the limited training data typical in medical imaging applications.

Model Configurations and Baselines. We compared Vision Mamba models against established CNN and Vision Transformer baselines. CNN baselines included ResNet-50 (He et al., 2016) and VGG-16 (Simonyan & Zisserman, 2015). ViT baselines comprised ViT-tiny-16, ViT-small-16, ViT-small-32, ViT-base-16, and ViT-base-32 (Dosovitskiy et al., 2021), where suffixes indicate model size and patch dimensions. Mamba-based models included Vim-small-16 (L. Zhu et al., 2024), VMamba-tiny, VMamba-small, and VMamba-base (Y. Liu et al., 2024).

Experimental Protocol. To ensure statistical robustness, we conducted each experiment across five training runs using different random seeds. Data splitting followed a 70/15/15 distribution for training, validation, and test sets. The validation set enabled early stopping to prevent overfitting, with the best-performing checkpoint selected for final evaluation on the test set. Statistical significance analysis employed paired t-tests comparing model predictions across all five folds, with significance threshold set at $p < 0.05$.

3.3.3 Results

Performance on Combined BUSI+B Dataset. Table 3.9 presents comprehensive results for the combined dataset. Mamba-based models demonstrated competitive or superior performance compared to CNN and ViT alternatives. VMamba-small achieved the highest AUC of 96.12 ± 0.75 , while VMamba-tiny achieved the highest accuracy of 89.36 ± 2.33 . Statistical significance analysis revealed that VMamba-tiny significantly outperformed VGG-16 ($p = 0.004$), ViT-tiny-16 ($p = 0.003$), ViT-small-32 ($p = 0.014$), and ViT-base-32 ($p = 0.022$). VMamba-base also significantly outperformed VGG-16 ($p = 0.037$) and ViT-tiny-16 ($p = 0.015$).

Performance on BUSI Dataset. Table 3.10 shows results for the three-class BUSI dataset. Vim-small-16 achieved the highest AUC of 95.63 ± 1.66 , while VMamba-base achieved the highest accuracy of 89.06 ± 3.72 . Statistical analysis revealed that VMamba-tiny significantly outperformed ResNet-50 ($p = 0.032$), VGG-16 ($p = 0.024$), and ViT-small-32 ($p = 0.002$). VMamba-base demonstrated the most comprehensive performance improvements, significantly outperforming ResNet-50 ($p = 0.015$), VGG-16 ($p = 0.006$), ViT-tiny-16 ($p = 0.018$), ViT-small-16 ($p = 0.022$), ViT-small-32

Table 3.9: Transfer learning results for BUSI+B dataset. AUC and accuracy values scaled 0-100, averaged over five runs.

Encoder Type	Encoder	Parameters (M)	AUC	ACC
CNN	ResNet-50	23.5	95.74 ± 1.42	87.66 ± 2.04
	VGG-16	134.3	94.25 ± 1.28	85.82 ± 1.49
ViT	tiny-16	5.5	94.19 ± 1.74	85.39 ± 1.93
	small-16	21.7	95.39 ± 0.54	87.23 ± 2.33
	small-32	22.5	93.85 ± 0.72	86.24 ± 1.65
	base-16	85.8	95.76 ± 0.77	88.51 ± 2.67
	base-32	87.5	95.51 ± 1.53	86.52 ± 3.23
Vim	small-16	25.4	95.84 ± 0.96	87.38 ± 3.22
VMamba	tiny	29.9	95.71 ± 1.01	89.36 ± 2.33
	small	49.4	96.12 ± 0.75	87.80 ± 2.78
	base	87.5	95.60 ± 0.79	88.51 ± 2.22

($p = 0.0001$), and ViT-base-32 ($p = 0.029$).

Table 3.10: Transfer learning results for BUSI dataset. AUC and accuracy values scaled 0-100, averaged over five runs.

Encoder Type	Encoder	Parameters (M)	AUC	ACC
CNN	ResNet-50	23.5	93.23 ± 1.93	85.64 ± 2.72
	VGG-16	134.3	93.69 ± 2.18	85.47 ± 4.59
ViT	tiny-16	5.5	93.52 ± 3.07	85.98 ± 4.48
	small-16	21.7	93.60 ± 4.04	86.50 ± 4.44
	small-32	22.5	93.59 ± 2.59	84.10 ± 3.81
	base-16	85.8	94.11 ± 2.12	87.18 ± 2.70
	base-32	87.5	94.10 ± 1.23	85.98 ± 1.39
Vim	small-16	25.4	95.63 ± 1.66	87.86 ± 2.72
VMamba	tiny	29.9	95.28 ± 1.89	88.55 ± 1.67
	small	49.4	94.48 ± 3.48	87.18 ± 4.15
	base	87.5	94.67 ± 2.53	89.06 ± 3.72

Performance on B Dataset. Table 3.11 presents results for the two-class B dataset, where VMamba-tiny achieved exceptional performance with AUC of 92.66 ± 9.07 and accuracy of 87.50 ± 12.08 , representing 1.98% AUC improvement and 5.0% accuracy improvement over the best non-Mamba model. Statistical analysis confirmed VMamba-tiny’s superiority, significantly outperforming ResNet-50 ($p = 0.011$), VGG-16 ($p = 0.028$), ViT-tiny-16 ($p = 0.004$), ViT-small-32 ($p = 0.012$), ViT-base-16 ($p = 0.003$), ViT-base-32 ($p = 0.007$), and even VMamba-base ($p = 0.034$).

Statistical Significance Summary. Across all datasets, Mamba-based models consistently

Table 3.11: Transfer learning results for B dataset. AUC and accuracy values scaled 0-100, averaged over five runs.

Encoder Type	Encoder	Parameters (M)	AUC	ACC
CNN	ResNet-50	23.5	90.05 ± 4.19	78.33 ± 5.53
	VGG-16	134.3	88.24 ± 7.10	80.00 ± 6.67
ViT	tiny-16	5.5	80.90 ± 9.98	77.50 ± 7.26
	small-16	21.7	90.68 ± 7.89	82.50 ± 6.67
	small-32	22.5	87.55 ± 8.89	79.17 ± 9.50
	base-16	85.8	88.32 ± 6.24	76.67 ± 7.26
	base-32	87.5	83.14 ± 12.82	77.50 ± 12.25
Vim	small-16	25.4	87.42 ± 9.83	84.17 ± 8.50
VMamba	tiny	29.9	92.66 ± 9.07	87.50 ± 12.08
	small	49.4	88.70 ± 9.30	83.33 ± 10.54
	base	87.5	92.19 ± 5.43	81.67 ± 6.24

demonstrated superior or equivalent performance with statistical significance. Crucially, no non-Mamba model significantly outperformed any Mamba-based model in our experiments, highlighting the robustness of the Vision Mamba approach for ultrasound classification tasks.

3.3.4 Discussion

Our evaluation demonstrates that Vision Mamba architectures offer significant advantages for breast ultrasound classification, with consistent performance improvements including 1.98% AUC and 5.0% accuracy gains on the B dataset. VMamba’s superior performance over Vim stems from its 2D selective scan mechanism enabling comprehensive spatial relationship modeling, while successful transfer learning from ImageNet demonstrates the architecture’s versatility despite significant domain gaps. The combination of long-range dependency modeling with inductive biases proves particularly valuable for ultrasound imaging where both local tissue characteristics and global anatomical context are essential.

The linear computational complexity makes these models ideal for point-of-care applications with memory constraints, positioning Vision Mamba as suitable for clinical deployment. However, study limitations include simplified statistical analysis due to multi-class AUC complexity and dataset imbalance. Future work should validate across larger, multi-institutional datasets to strengthen generalization claims for clinical adoption.

3.4 Concluding Discussions

Our investigation of Vision Mamba architectures across histopathology and breast ultrasound imaging reveals that efficient representation learning in medical domains requires architectures that align with domain-specific processing patterns while maintaining superior performance.

The comparative analysis reveals distinct performance advantages that manifest differently across domains. In computational pathology, Vim4Path achieved substantial improvements, with Vim-tiny demonstrating an 8.21 AUC point gain over ViT-tiny (95.81 vs 87.60) with comparable parameter counts. The improvements proved particularly pronounced at smaller model scales, where Vim-tiny-plus (13M parameters) outperformed the larger ViT-small (22M parameters), demonstrating parameter efficiency advantages. In ultrasound classification, VMamba-tiny achieved 1.98% AUC and 5.0% accuracy improvements over the best non-Mamba baseline, with statistical significance analysis confirming that Mamba-based models were never significantly outperformed by traditional architectures.

The architectural differences between Vim and VMamba demonstrate important design considerations. Vim’s bidirectional scanning proved highly effective for histopathology’s patch-based processing, while VMamba’s 2D selective scan mechanism demonstrated superior performance in ultrasound applications. This suggests that architectural variants within the Vision Mamba family offer domain-specific advantages based on imaging characteristics.

Although we have not yet evaluated Mamba-based models on high-resolution images, one of their primary strengths is memory-efficient processing of long contexts (Gu & Dao, 2024). In this study we therefore concentrated on standard-resolution images to demonstrate the models’ effectiveness and to lay the groundwork for future work with larger image sizes.

The explainability analysis revealed that Vim learns representations that align more closely with domain expertise. In histopathology, Vim attention patterns focused on specific biological features used by pathologists for diagnostic validation, while Vision Transformers adopted more holistic approaches. This alignment suggests that sequential processing constraints can improve feature specificity and interpretability without sacrificing performance, creating a foundation for trustworthy clinical decision support systems where efficient architectures simultaneously enhance

both computational performance and alignment with expert knowledge. While our GradCAM-based analysis provided initial insights into Vision Mamba’s interpretability advantages, these findings motivated a deeper investigation into how modern foundation models in computational pathology learn and represent domain knowledge.

While our GradCAM analysis provided initial insights into Vision Mamba’s interpretability advantages, these findings motivated a deeper investigation into how modern foundation models in computational pathology learn and represent domain knowledge. GradCAM and similar saliency methods have well known limitations: they can highlight spurious correlations, vary dramatically with small input perturbations, and fundamentally only show where models attend, not what concepts they’ve learned. In medical domains where understanding the actual features models recognize is essential for clinical trust, we need more rigorous approaches.

Sparse Autoencoders (SAEs) offer a principled solution by learning to decompose model activations into sparse, interpretable components. Unlike saliency methods that provide post hoc visualizations, SAEs directly access and disentangle the internal representations models use for decision making. By enforcing sparsity, SAEs aim to isolate monosemantic features where individual latent dimensions correspond to distinct, human understandable concepts. This mechanistic approach has shown promise in language models for revealing interpretable features, making it a natural choice for understanding what medical concepts pathology models encode.

The next chapter presents a comprehensive interpretability analysis using Sparse Autoencoders to compare the learned representations of current pathology foundation models, moving beyond simple gradient based explanations to more rigorous mechanistic interpretability approaches.

Chapter 4

Sparse Autoencoders and interpretability for histopathology images

The interpretability insights from Chapter 3, while encouraging, exposed the limitations of gradient-based analysis methods. Although GradCAM ([Selvaraju et al., 2017](#)) suggested that Vision Mamba attends to diagnostically relevant features, such saliency methods cannot reveal what concepts models actually encode. They merely highlight spatial regions without explaining the learned representations driving those attention patterns. This limitation becomes critical when considering the broader landscape of pathology foundation models.

As computational pathology rapidly develops diverse foundation models including Phikon ([Filiot et al., 2023](#)), Quilt ([Ikezogwo et al., 2023](#)), UNI ([R. J. Chen et al., 2024](#)), Virchow ([Vorontsov et al., 2023](#)), and CHIEF ([X. Wang et al., 2024](#)), each trained on millions of images with different architectures and objectives, a fundamental question emerges: what medical concepts have these models learned? Understanding their internal representations is not merely academic curiosity; it directly impacts clinical trust, regulatory approval, and our ability to identify potential biases or failure modes before deployment.

This chapter applies Sparse Autoencoders (Bricken et al., 2023) to decode the internal representations of pathology foundation models, seeking to move beyond surface level visualizations to mechanistic understanding. However, our investigation reveals an unexpected challenge: current interpretability methods create incompatible, model specific representations that scale exponentially with the number of models, threatening to bottleneck our understanding of rapidly advancing medical AI.

4.1 Sparse Autoencoder Framework for Model Interpretability

Sparse Autoencoders (SAEs) have emerged as a powerful tool for understanding what neural networks learn by decomposing complex model activations into interpretable, sparse representations (Bricken et al., 2023). Unlike gradient-based attribution methods that provide post-hoc explanations, SAEs learn a new basis that ideally makes individual dimensions correspond to distinct, human-understandable concepts.

4.1.1 Mathematical Formulation

Let $I \in \mathbb{R}^{H \times W \times C}$ denote an input image processed through a pre-trained foundation model $F(\cdot; \theta_F)$ to extract features $x = F(I; \theta_F)$ where $x \in \mathbb{R}^K$. The SAE operates on these extracted features as follows:

$$\bar{x} = x - b_{tied} \quad (\text{Pre-Encoder Bias}) \quad (1)$$

$$h = \text{ReLU}(W_e \bar{x} + b_e) \quad (\text{Encoder}) \quad (2)$$

$$\tilde{x} = W_d h \quad (\text{Decoder}) \quad (3)$$

$$\hat{x} = \tilde{x} + b_{tied} \quad (\text{Post-Decoder Bias}) \quad (4)$$

Where $b_{tied} \in \mathbb{R}^K$ is a tied bias parameter, $W_e \in \mathbb{R}^{L \times K}$ is the encoder weight matrix, $b_e \in \mathbb{R}^L$ is the encoder bias, $h \in \mathbb{R}^L$ is the latent activation vector, and $W_d \in \mathbb{R}^{K \times L}$ is the unit norm decoder weight matrix.

The SAE is trained using a combination of reconstruction error and sparsity regularization:

$$L_{total} = \frac{1}{n} \sum_{i=1}^n (\|x_i - \hat{x}_i\|_2^2 + \lambda_{sparse} \cdot \|h_i\|_1) \quad (5)$$

The tied bias mechanism allows the model to focus on learning variations rather than encoding constant offsets, while the unit norm constraint on decoder weights prevents artificial scaling to reduce sparsity loss (Bricken et al., 2023). The resulting sparse latent representations h ideally contain monosemantic dimensions where each neuron responds to a single, distinct concept rather than multiple entangled features.

4.2 SAE Analysis of Pathology Foundation Models

We applied SAEs to analyze pathology foundation models, focusing on Phikon (vision-only) and Quilt (vision-language) as representative examples. Each model was analyzed using patches extracted from histopathology slides, with SAEs trained to decompose the models’ learned representations into interpretable sparse components.

4.2.1 The Cross-Model Comparison Challenge

Figure 4.1 illustrates a fundamental challenge that emerged from this analysis. The figure shows representative SAE dimension activations, comparing Phikon (vision-only model) with Quilt (vision-language model). We selected both strongly aligned and weakly aligned features to demonstrate the range of interpretability challenges across different foundation models.

The analysis reveals several critical issues with individual model SAE analysis. In Quilt, SAE dimension 340 predominantly highlights normal colon mucosa, demonstrating good monosemantic behavior. However, dimension 343 activates for both loose stroma and dense fibrous stroma, indicating mixed feature encoding. Similarly, in Phikon, SAE dimension 252 associates with normal breast lobules, while dimension 328 responds to both myxoid stroma and immune-infiltrated stroma, suggesting polysemantic behavior.

More fundamentally, there is no systematic way to identify corresponding concepts across the

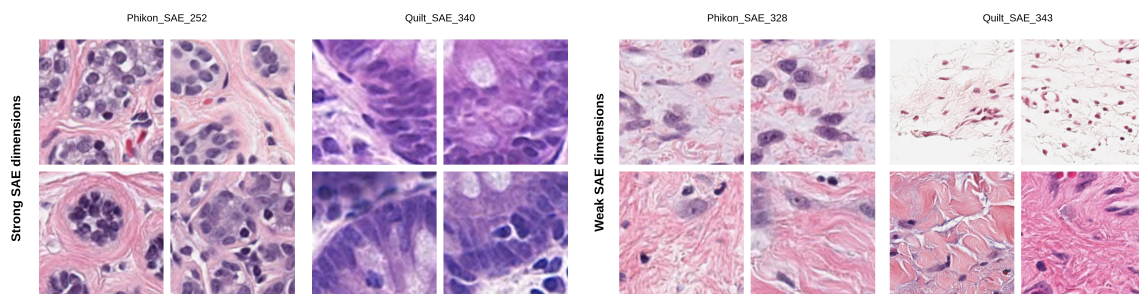


Figure 4.1: Representative samples from SAE dimension analysis for Phikon and Quilt. We show both strong SAE dimensions with high consistency as well as dimensions with lower consistency in their attributes.

two models. Each model’s SAE produces its own isolated latent space with no inherent correspondence between dimensions. Understanding what concepts different models have learned requires analyzing each model’s SAE output individually and manually identifying potential correspondences across architectures.

4.3 The Scalability Problem in Cross-Model Interpretability

The results from our SAE analysis expose a fundamental limitation in current interpretability paradigms. While benchmarking model performance scales linearly, we run each model once to obtain comparable metrics, interpretability analysis scales exponentially with both the number of models and the complexity of comparisons required.

4.3.1 Manual Expert Analysis Bottleneck

In computational pathology, validating SAE interpretations requires pathologist expertise to determine whether identified features correspond to meaningful biological concepts. This creates several compounding challenges:

- **Model-Specific Analysis:** Each foundation model requires separate SAE training, hyperparameter tuning, and expert evaluation
- **Concept Mapping:** Identifying corresponding concepts across models demands manual comparison of hundreds or thousands of SAE dimensions

- **Expert Time Constraints:** Pathologist time is expensive and limited, creating a bottleneck that cannot scale with rapid model development
- **Inconsistent Latent Spaces:** Without systematic alignment, comparing model representations becomes an intractable manual process

This bottleneck becomes prohibitive as the field develops increasingly sophisticated foundation models. Analyzing five pathology models would require expert evaluation of five separate SAE outputs, with no systematic framework for identifying shared or distinct concepts across architectures.

4.3.2 The Need for Unified Interpretability

The computational pathology field exemplifies a broader challenge facing interpretability research: we can develop and benchmark new models faster than we can understand what they learn. While performance metrics enable rapid comparison across dozens of architectures, interpretability analysis remains trapped in a paradigm of individual model examination that cannot keep pace with model development.

This limitation is particularly important in specialized domains like pathology, where interpretability insights require domain expertise that is both expensive and scarce. The current approach of analyzing each model separately not only wastes expert time but also prevents systematic understanding of how different architectural choices and training objectives affect learned representations.

4.4 Toward Cross-Model Interpretability

The challenges identified in this chapter highlight the critical need for interpretability frameworks that can operate across multiple models simultaneously. Rather than requiring separate analysis of each architecture, we need approaches that learn unified representations enabling direct comparison of concepts across different foundation models.

Such a framework would transform interpretability analysis from an exponentially scaling manual process into a systematic, scalable approach that can keep pace with rapid model development. By learning shared concept representations, experts could analyze what different models have

learned through a single unified interface, dramatically reducing the time and expertise required for cross-model interpretability.

Chapter 5 presents SPARC (**S**pars**e** Autoencoders for Aligned **R**epresentation of **C**oncepts) (Nasiri-Sarvi, Rivaz, & Hosseini, 2025), our solution to these scalability challenges. SPARC learns a single, shared latent space that works across multiple models simultaneously, enabling direct comparison of how different architectures represent identical concepts without requiring model-specific analysis or manual alignment procedures.

Chapter 5

Scalable Interpretability and Representation Engineering with SPARC

Chapter 4 demonstrated the fundamental scalability problem facing interpretability research: while performance benchmarking scales linearly with the number of models, interpretability analysis scales exponentially, creating an unsustainable bottleneck as model development accelerates. Our sparse autoencoder analysis of pathology foundation models revealed that each model produces its own isolated latent space with no inherent correspondence between dimensions, forcing experts to perform labor-intensive analyses on each model separately and manually align findings across architectures.

This chapter presents SPARC (**S**parsE Autoencoders for **A**ligned **R**epresentation of **C**oncepts) (Nasiri-Sarvi et al., 2025), our solution to the cross-model interpretability challenges identified in the previous chapter. SPARC learns a single, unified latent space that works across multiple models simultaneously, enabling direct comparison of how different architectures represent identical concepts without requiring model-specific analysis or manual alignment procedures.

5.1 Introduction: From Problem to Solution

The cross-model comparison challenge illustrated in Figure 4.1 demonstrates the broader scalability issues in interpretability research. When we applied traditional sparse autoencoders to pathology foundation models like Phikon (Filiot et al., 2023) and Quilt (Ikezogwo et al., 2023), we discovered that understanding what concepts different models have learned requires analyzing each model’s SAE output individually and manually identifying potential correspondences across architectures. This approach cannot keep pace with the rapid development of foundation models.

The practical implications are substantial: analyzing five pathology foundation models using traditional SAE approaches would require training five separate autoencoders, conducting five independent expert evaluations, and manually comparing hundreds or thousands of latent dimensions to identify corresponding concepts. As specialized domains like computational pathology develop increasingly sophisticated foundation models, from Phikon (Filiot et al., 2023) and UNI (R. J. Chen et al., 2024) to Virchow (Vorontsov et al., 2023) and CHIEF (X. Wang et al., 2024), this exponential scaling becomes prohibitively expensive.

The challenge extends beyond computational pathology to any domain where interpretability insights require specialized knowledge. Recent advances in vision-language models like CLIP (Radford et al., 2021), vision-only models like DINO (Oquab et al., 2024), and domain-specific foundation models across medical imaging, autonomous driving, and scientific domains all face the same limitation: we can develop and benchmark new models faster than we can understand what they learn.

This chapter addresses these challenges through SPARC, a method that transforms interpretability analysis of multiple models from an exponentially scaling manual process into a systematic, scalable approach. By learning shared concept representations, experts can analyze what different models have learned through a single unified interface, substantially reducing the time and expertise required for cross-model interpretability while enabling new capabilities like cross-modal attribution and systematic concept comparison across architectures.

5.2 SPARC Method: Shared Sparse Representations Across Models

Traditional sparse autoencoders, as demonstrated in Chapter 4, produce incompatible concept spaces when applied to different models. SPARC addresses this limitation through two key innovations that enable concept alignment across heterogeneous models and modalities: Global TopK sparsity for structural alignment and cross-reconstruction loss for semantic alignment.

5.2.1 Problem Formulation

We consider multiple (M) distinct streams of information, indexed by the set $\mathcal{S} = \{s_1, s_2, \dots, s_M\}$. Unlike the isolated analysis approach that created the scalability problems identified in Chapter 4, SPARC processes data samples with multiple related representations of the same underlying entity simultaneously. For instance, features from an image can be extracted using both CLIP-image (Radford et al., 2021) and DINO (Oquab et al., 2024) models, or image-caption pairs can be processed through DINO and CLIP-text encoders.

For each sample, we obtain a set of corresponding feature vectors $\{\mathbf{x}^s\}_{s \in \mathcal{S}}$, where each $\mathbf{x}^s \in \mathbb{R}^{d_s}$ is produced by stream s 's feature extraction process with dimensionality d_s . These streams represent heterogeneous processing pathways that may use different architectures (CNN vs. Transformer), training objectives (contrastive vs. supervised), or process different modalities associated with the input (vision vs. language). Consequently, our framework must accommodate varying input feature dimensionalities ($d_s \neq d_t$ for $s \neq t$).

Given different feature representations $\{\mathbf{x}^s\}_{s \in \mathcal{S}}$ of the same underlying data sample, SPARC maps these inputs into concept-aligned sparse latent representations $\{\mathbf{z}^s\}_{s \in \mathcal{S}}$. These latent representations share a common L -dimensional space where $\mathbf{z}^s \in \mathbb{R}^L$, maintain sparsity $\|\mathbf{z}^s\|_0 \ll L$, and crucially activate the same semantic dimensions across different streams when processing identical underlying content.

The design is guided by three key objectives that directly address the limitations identified in Chapter 4:

- (1) **Faithfulness:** The model aims to accurately reconstruct the original features for each stream

from its latent representation, ensuring the learned interpretable model captures essential information.

- (2) **Interpretability:** Enforced sparsity promotes monosemanticity, where individual latent dimensions ideally capture distinct, human-understandable concepts, reducing semantic entanglement.
- (3) **Concept Alignment:** The interpretation of each latent dimension should remain consistent across all input streams, enabling direct comparison without manual alignment procedures.

5.2.2 Architecture Design

Figure 5.1 illustrates the complete SPARC architecture, which addresses the cross-model comparison challenges through two key innovations that distinguish it from traditional sparse autoencoders and prior cross-model approaches like Universal Sparse Autoencoders (USAE) (Thasarathan et al., 2025).

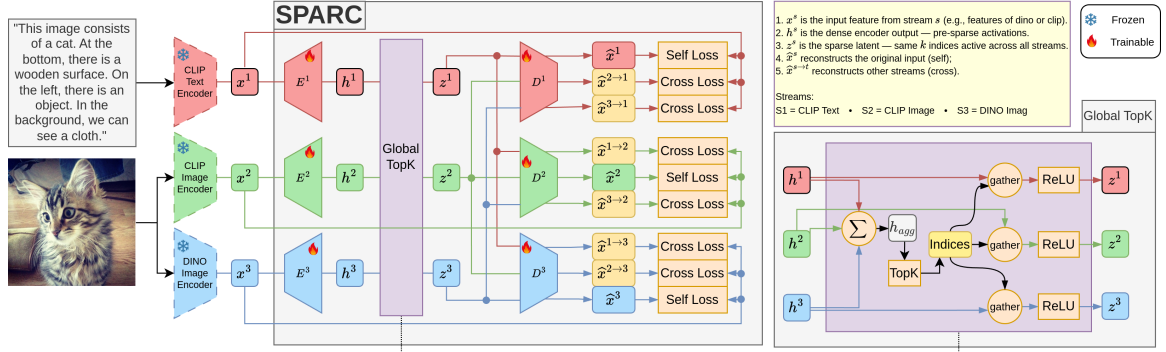


Figure 5.1: Detailed architecture of the SPARC model as well as the Global TopK mechanism.

Stream Encoders. Each input feature vector $\mathbf{x}^s \in \mathbb{R}^{d_s}$ is processed by its corresponding stream-specific encoder $E_s : \mathbb{R}^{d_s} \rightarrow \mathbb{R}^L$. The encoder performs an affine transformation to map the input features to the L -dimensional latent space, producing pre-activation logits \mathbf{h}^s :

$$\mathbf{h}^s = E_s(\mathbf{x}^s) = \mathbf{W}_E^s(\mathbf{x}^s - \mathbf{b}_{\text{pre}}^s) + \mathbf{b}_{\text{lat}}^s \quad (6)$$

where $\mathbf{b}_{\text{pre}}^s \in \mathbb{R}^{d_s}$ is a learnable pre-bias, $\mathbf{W}_E^s \in \mathbb{R}^{L \times d_s}$ are the encoder weights, and $\mathbf{b}_{\text{lat}}^s \in \mathbb{R}^L$ is a learnable latent bias.

Global TopK Sparse Activation. A core innovation addressing the alignment challenges identified in Chapter 4 is SPARC’s Global TopK mechanism. Rather than applying sparsity independently to each stream’s logits \mathbf{h}^s (which would reproduce the incompatible latent spaces problem), SPARC enforces shared feature activation across streams.

Logits are first aggregated across all streams:

$$\mathbf{h}_{\text{agg}} = \sum_{s \in \mathcal{S}} \mathbf{h}^s \quad (7)$$

The top- k indices are then selected using this aggregated logit:

$$\mathcal{I}_{\text{global}} = \text{TopK}(\mathbf{h}_{\text{agg}}, k) \quad (8)$$

This shared index set $\mathcal{I}_{\text{global}}$ is used to construct sparse latent representations for each stream:

$$\mathbf{z}^s = \text{ReLU}(\text{gather}(\mathbf{h}^s, \mathcal{I}_{\text{global}})) \quad (9)$$

where $\text{gather}(\mathbf{h}^s, \mathcal{I}_{\text{global}})$ keeps values from \mathbf{h}^s at positions in $\mathcal{I}_{\text{global}}$ and zeros elsewhere. This ensures both sparsity ($\|\mathbf{z}^s\|_0 \leq k$) and concept alignment by forcing identical latent features to activate across all streams for the same underlying data sample.

Stream Decoders. Each stream-specific decoder $D_s : \mathbb{R}^L \rightarrow \mathbb{R}^{d_s}$ maps the sparse latent representation \mathbf{z}^s back to the original feature space:

$$\hat{\mathbf{x}}^s = D_s(\mathbf{z}^s) = \mathbf{W}_D^s \mathbf{z}^s + \mathbf{b}_{\text{pre}}^s \quad (10)$$

5.2.3 Training Objective

SPARC’s training objective addresses the semantic alignment challenge through a combination of self-reconstruction and cross-reconstruction losses. While Global TopK provides a hard structural constraint ensuring the same set of TopK neurons activate across streams, the cross-reconstruction

loss provides a soft semantic constraint that encourages each individual activated neuron to encode consistent concepts across streams through optimization.

The optimization relies on two fundamental reconstruction pathways:

Self-reconstruction. Stream s reconstructs its own input using its latent representation:

$$\hat{\mathbf{x}}^s = D_s(\mathbf{z}^s) \quad (11)$$

Cross-reconstruction. Stream t 's decoder reconstructs its input using stream s 's latent representation:

$$\hat{\mathbf{x}}^{s \rightarrow t} = D_t(\mathbf{z}^s) = \mathbf{W}_D^t \mathbf{z}^s + \mathbf{b}_{\text{pre}}^t \quad (s \neq t) \quad (12)$$

To handle varying feature scales across streams, we use Normalized Mean Squared Error (NMSE):

$$\mathcal{L}_{\text{NMSE}}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2}{\|\mathbf{x}\|_2^2} \quad (13)$$

The final training objective combines self-reconstruction and cross-reconstruction losses:

$$\mathcal{L}_{\text{total}} = \underbrace{\sum_{s \in \mathcal{S}} \mathcal{L}_{\text{NMSE}}(\mathbf{x}^s, \hat{\mathbf{x}}^s)}_{\mathcal{L}_{\text{self}}} + \lambda \underbrace{\sum_{\substack{s, t \in \mathcal{S} \\ s \neq t}} \mathcal{L}_{\text{NMSE}}(\mathbf{x}^t, \hat{\mathbf{x}}^{s \rightarrow t})}_{\mathcal{L}_{\text{cross}}} \quad (14)$$

The hyperparameter $\lambda \geq 0$ balances faithfulness (through self-reconstruction) with concept alignment (through cross-reconstruction). Together, this training objective and the Global TopK mechanism address all three objectives: faithfulness through self-reconstruction, interpretability through enforced sparsity, and concept alignment through both shared activation patterns and cross-stream transferability.

5.3 Experimental Setup

Our experimental evaluation addresses the scalability challenges identified in Chapter 4 by demonstrating SPARC's ability to create unified interpretable representations across heterogeneous models and modalities. We focus on vision and vision-language models to validate cross-modal

alignment capabilities while providing systematic comparison against traditional approaches.

5.3.1 Problem Setup and Model Streams

We evaluate SPARC on $M=3$ distinct information streams representing different architectural paradigms and training objectives. Each stream processes the same underlying data sample but produces feature representations with different dimensionalities and semantic structures:

- **DINO** (Oquab et al., 2024): Self-supervised vision-only model (1024-dimensional features) trained with knowledge distillation
- **CLIP-Image** (Radford et al., 2021): Vision encoder from CLIP model (768-dimensional features) trained with contrastive vision-language objectives
- **CLIP-Text** (Radford et al., 2021): Text encoder from CLIP model (768-dimensional features) processing image captions

This heterogeneous combination enables evaluation of both cross-model alignment (DINO vs CLIP-Image, both processing visual information through different architectures and training objectives) and cross-modal alignment (vision streams vs CLIP-Text language stream). The varying input dimensionalities test SPARC’s ability to create shared representations across fundamentally different feature spaces.

5.3.2 Architecture and Training Configuration

SPARC maps all input streams into a shared $L=8,192$ dimensional latent space with sparsity level $k=64$, meaning only 0.78% of latent dimensions activate for any given sample. This high sparsity promotes monosemanticity while the large latent space provides sufficient capacity for diverse concept representation. Our training objective combines self-reconstruction and cross-reconstruction losses with equal weighting ($\lambda = 1.0$). Self-reconstruction requires each stream to faithfully reconstruct its own features from its latent representation, ensuring information preservation. Cross-reconstruction requires each stream’s latent representation to reconstruct features from other streams, creating optimization pressure toward shared semantic understanding. We use

Normalized Mean Squared Error (NMSE) to handle the varying feature scales across streams. The Global TopK mechanism aggregates logits across all streams before selecting the top-k indices, ensuring identical activation patterns across streams for the same underlying content. This structural constraint addresses the core problem identified in Chapter 4 where traditional SAEs produce incompatible latent spaces.

5.3.3 Datasets and Experimental Scope

Open Images V7 (Kuznetsova et al., 2020) serves as our primary dataset, using the dense-annotated subset containing 1.7M training images with hierarchical taxonomic labels. We evaluate concept alignment across 432 binary classification tasks (filtered from 601 total tasks to ensure ≥ 50 positive examples per task). The hierarchical taxonomy structure proves essential for measuring semantic alignment at different granularities. For instance, when one stream activates for "tiger" and another for "leopard," both correctly align at the "carnivore" level. **MS-COCO 2017** (Lin et al., 2014) is mainly used for downstream applications, testing whether aligned representations learned on Open Images generalize to semantic segmentation and cross-modal retrieval tasks.

5.3.4 Baseline Comparisons and Ablation Design

We design our baseline comparisons to systematically validate each component of SPARC:

Local TopK ($\lambda = 0$): Traditional sparse autoencoder approach where each stream independently selects top-k activations with only self-reconstruction loss. This represents the standard approach that created incompatible latent spaces in Chapter 4.

Local TopK + Cross-loss ($\lambda = 1$): Tests whether semantic alignment can be achieved through cross-reconstruction loss alone, without structural activation constraints.

Global TopK + No Cross-loss ($\lambda = 0$): Tests whether structural alignment through shared activation patterns suffices without explicit semantic supervision.

SPARC (Global TopK + Cross-loss, $\lambda = 1$): Our complete approach combining both structural and semantic alignment mechanisms.

This 2x2 ablation design isolates the contribution of each innovation and validates our hypothesis that both components are necessary for effective concept alignment.

5.3.5 Evaluation Framework

Our evaluation addresses five critical aspects of the cross-model interpretability challenge:

Activation Pattern Analysis: We quantify neuron activation patterns across all 8,192 latent dimensions, measuring the percentage of neurons that are alive across all streams, dead across all streams, or exhibit mixed patterns (active in some streams but not others). This directly tests whether SPARC eliminates the incompatible latent spaces problem.

Semantic Concept Alignment: We measure concept consistency using hierarchical Jaccard similarity on Open Images taxonomy. For each latent dimension and stream, we identify the top-50 activating samples and compute label overlap across stream pairs at different taxonomy depths (from fine-grained leaf concepts to broad categories). This tests whether structurally aligned activations represent semantically equivalent concepts.

Monosemantic Concept Recovery: Following established SAE evaluation protocols ([Gao et al., 2025](#); [Gurnee et al., 2023](#)), we assess whether individual latent dimensions encode recognizable semantic features using 1D logistic probes. For each of the 432 binary classification tasks in Open Images, we train simple logistic regression classifiers that use single latent dimension activations to predict concept presence. We select the top-20 most frequently activated latent dimensions per task as candidates and report the best-performing dimension’s cross-entropy loss. This evaluation tests whether SPARC’s aligned representations maintain interpretability, scores below the random baseline (0.693 for balanced binary classification) indicate that individual latent dimensions capture meaningful, linearly separable concepts.

Downstream Application Validation: We evaluate practical utility through two applications enabled by aligned representations: (1) weakly supervised semantic segmentation on MS-COCO using SPARC latents as scalar targets for gradient-based attribution, and (2) cross-modal retrieval in the aligned latent space. These tasks validate that concept alignment translates to useful capabilities rather than merely improving abstract metrics.

Reconstruction Quality Assessment: We monitor both self-reconstruction and cross-reconstruction NMSE to ensure that alignment improvements do not come at the cost of representational fidelity. This addresses the fundamental trade-off between individual model performance and cross-model

interpretability.

This comprehensive evaluation enables systematic assessment of whether SPARC successfully transforms the exponentially scaling interpretability analysis problem into a scalable, unified approach for understanding concept representations across diverse AI architectures.

5.4 Results and Analysis

Our experimental results demonstrate that SPARC successfully addresses the cross-model interpretability scalability challenges identified in Chapter 4. The combination of Global TopK and cross-reconstruction loss produces substantial improvements in concept alignment while enabling new capabilities for systematic cross-model analysis.

5.4.1 Latent Activation Alignment

Figure 5.2 shows neuron 6463’s top-activating images across four training configurations. In Local TopK with $\lambda = 1$ (top right), CLIP-text stream shows no activations (dead neuron) while DINO and CLIP-image streams activate on different image types. In contrast, Global TopK with $\lambda = 1$ (bottom right) shows consistent activations across all three streams for the same object type.

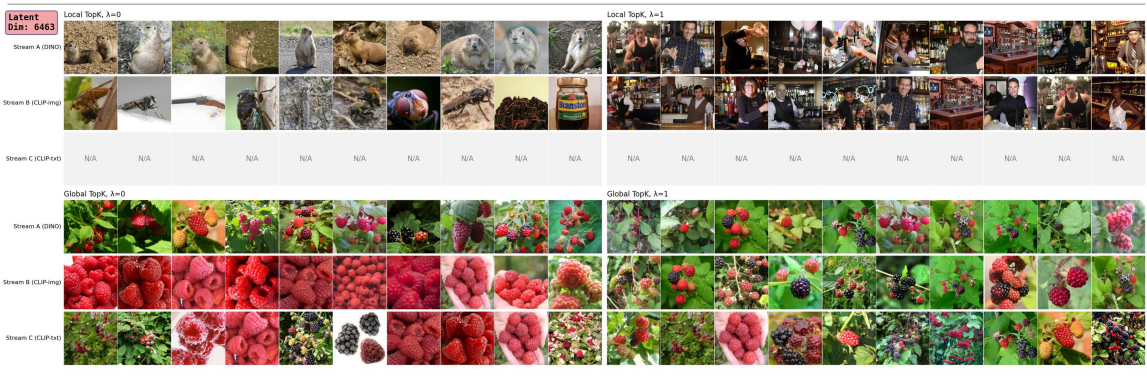


Figure 5.2: Top-activating samples for the latent dimension 6463 across three streams (DINO, CLIP-img, CLIP-txt) under four SPARC configurations. Each row shows top-10 images that activate the latent. The CLIP-text stream shows no activations under Local TopK with $\lambda = 1$ due to a dead neuron.

We find this pattern to be common where without Global TopK, one or two streams have dead latents while others remain active. We quantified this pattern across all 8,192 latent dimensions (Table 5.1). Local TopK with $\lambda = 1$ produces 48.8% mixed activation patterns where only 2/3 streams are active, creating alignment failures. Global TopK with $\lambda = 1$ achieves 84.4% all-alive neurons with consistent cross-stream activation and 0% cases of partial activation. This means with Global TopK with $\lambda = 1$, all neurons are either active across all streams or dead across all streams. We also report per-stream percentages of dead neurons, showing that with Local TopK with $\lambda = 1$, almost half of the CLIP-text neurons are dead. With Global TopK with $\lambda = 1$, dead neurons are distributed equally across all streams.

Table 5.1: Neuron activation patterns and stream-specific dead neuron rates across 8192 latent dimensions. Mixed patterns indicate partial cross-stream alignment where only 1/3 or 2/3 of the streams are active for the same latent. CI = CLIP-image, CT = CLIP-text, D = DINO.

TopK	λ	Activation Patterns				Dead Neurons		
		(%)				(%)		
		All Dead	1/3	2/3	All Alive	CI	CT	D
Global	1.0	15.6	0.0	0.0	84.4	15.6	15.6	15.6
Global	0.0	1.6	0.0	0.2	98.2	1.6	1.8	1.6
Local	1.0	0.0	7.6	48.8	43.6	16.5	45.0	2.4
Local	0.0	0.0	0.0	14.3	85.7	0.3	14.1	0.0

The results reveal dramatic differences in activation patterns. Local TopK with cross-loss produces 48.8% mixed activation patterns where only 2/3 streams are active, creating the alignment failures that motivated this work. In contrast, Global TopK with cross-loss achieves 84.4% all-alive neurons with consistent cross-stream activation and completely eliminates mixed activation patterns (0.0% cases of partial activation).

This means SPARC ensures that all neurons are either active across all streams or dead across all streams, addressing the incompatible latent spaces problem. The equal distribution of dead neurons across streams (15.6% for each) confirms that the Global constraint works as intended, while Local TopK shows highly unbalanced dead neuron rates, particularly for CLIP-text (45.0% dead neurons).

5.4.2 Quantitative Concept Alignment

We evaluate whether individual latent dimensions consistently represent the same concepts across streams using hierarchical Jaccard similarity on Open Images taxonomy. This addresses cases where related concepts like "tiger" and "leopard" should align at higher semantic levels (both mapping to "carnivore").

For each latent and stream, we identify top-50 activating samples and measure alignment using generalized Jaccard similarity between label distributions across stream pairs at different taxonomy depths.

Table 5.2 shows Global TopK with cross-loss ($\lambda = 1$) achieves 0.80 Jaccard similarity at finest granularity, more than tripling alignment compared to Local TopK (0.26). The results demonstrate that both Global TopK and cross-reconstruction loss are essential for achieving concept alignment, with consistent improvements across all semantic granularities from broad categories to fine-grained concepts.

Table 5.2: Mean Jaccard similarity across Open Images taxonomy depths, grouped by TopK type and cross-loss weight λ . Depth 0 corresponds to full collapse into the root category (`Entity`), while depth 5 corresponds to no collapse (leaf-level granularity).

TopK Type	λ	Jaccard @ Depth					
		0 (root)	1	2	3	4	5 (leaf)
Global	1	0.8118	0.8056	0.8032	0.8020	0.8018	0.8018
Global	0	0.8054	0.7633	0.7451	0.7359	0.7344	0.7344
Local	1	0.3238	0.2866	0.2706	0.2615	0.2600	0.2599
Local	0	0.3458	0.2031	0.1783	0.1666	0.1652	0.1651

5.4.3 Monosemantic Concept Recovery

To assess whether SPARC maintains interpretability while achieving cross-model alignment, we evaluate monosemantic concept recovery using 1D logistic probes following established SAE protocols (Gao et al., 2025; Gurnee et al., 2023). We use 432 binary classification tasks from Open Images (filtered from 601 total tasks requiring ≥ 50 positive samples) to test whether individual latent dimensions encode distinct, linearly separable semantic concepts. For each task, we train 1D logistic regression probes that predict concept presence using single latent dimension activations,

minimizing binary cross-entropy loss:

$$\min_{i,w,b} \mathbb{E} [y \log \sigma(wz_i + b) + (1 - y) \log(1 - \sigma(wz_i + b))] \quad (15)$$

where z_i is the i -th latent dimension activation and σ is the sigmoid function. We evaluate the top-20 most activated dimensions per task and report the best-performing dimension’s loss, averaged across all tasks.

Table 5.3: Mean probe loss (lower is better) across 432 Open Images binary classification tasks.

TopK Type	λ	CLIP-Img	CLIP-Txt	DINO
Global	1.0	0.5355	0.5646	0.5409
Global	0.0	0.5336	0.4942	0.5194
Local	1.0	0.4990	0.5363	0.5170
Local	0.0	0.5238	0.4904	0.5265

Table 5.3 presents the mean probe loss results across different SPARC configurations. All configurations achieve probe losses substantially below the random baseline of 0.693, confirming that individual latent dimensions successfully encode linearly separable semantic concepts. Local TopK configurations generally achieve slightly lower losses than Global TopK, and configurations without cross-reconstruction loss ($\lambda = 0$) tend to produce better individual concept recovery.

These results reveal a trade-off between concept alignment and individual concept clarity. While SPARC’s alignment mechanisms successfully achieve cross-model interpretability, they come at a modest cost to individual concept precision. However, all configurations maintain interpretability well above random performance, preserving the fundamental monosemantic properties essential for sparse autoencoder utility.

5.4.4 Downstream Applications: Semantic Segmentation

We test whether SPARC’s concept-aligned latents can enable weakly supervised semantic segmentation by serving as scalar targets for gradient-based attribution methods. Gradient-based attribution techniques like GradCAM (Selvaraju et al., 2017) and relevancy maps (Chefer et al., 2021b) require scalar targets to compute spatial attributions via $\nabla A = \frac{\partial \text{target}}{\partial A}$.

SPARC enables cross-modal dot products like $\mathbf{z}^{\text{dino}} \cdot \mathbf{z}^{\text{clip_txt}}$ that would be meaningless with separate SAEs due to incompatible latent spaces. We evaluate two SPARC approaches against established baselines on MS-COCO validation set (note that SPARC was trained on Open Images, providing out-of-distribution evaluation).

Cross-Modal-Sim computes cross-modal similarities in the aligned latent space: $\mathbf{z}^{\text{clip_txt}} \cdot \mathbf{z}^{\text{dino}}$ for SPARC DINO and $\mathbf{z}^{\text{clip_txt}} \cdot \mathbf{z}^{\text{clip_img}}$ for SPARC CLIP.

Concept-Latent-Sum uses concept-specific latent selections $\sum_{j \in \mathcal{J}} z_j^s$ where \mathcal{J} contains latents that activate frequently (≥ 50 times) for the target class. We compare against DETR attention maps and CLIP similarity baselines using the evaluation protocol of [Chefer et al. \(2021b\)](#).

Figure 5.3 illustrates SPARC’s key capability: individual concept-specific latents produce coherent spatial heatmaps across DINO and CLIP vision encoders while generating meaningful text attribution scores in CLIP’s text encoder.

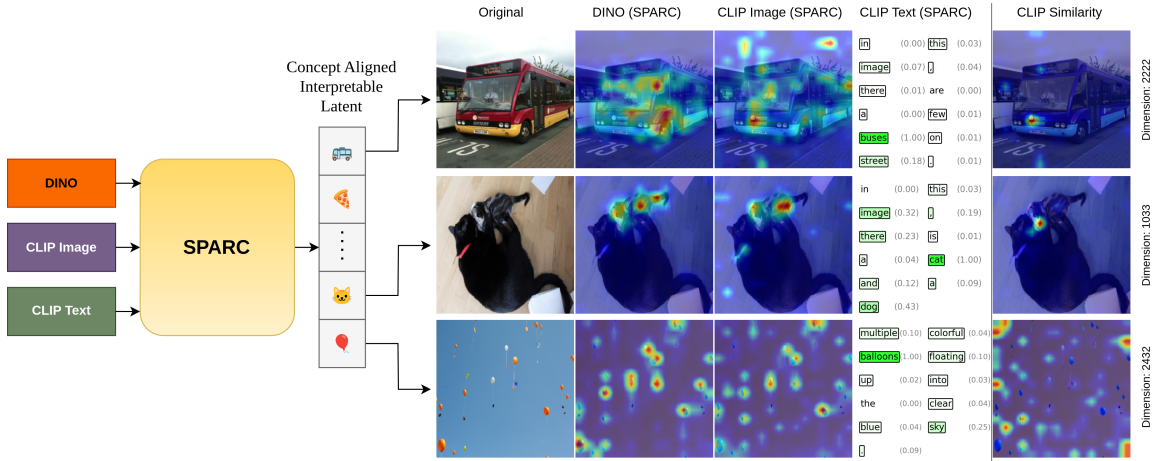


Figure 5.3: SPARC enables consistent concept visualization across models and modalities using shared latent dimensions. The figure demonstrates how individual concept-specific latents (bus, cat, balloons) produce coherent spatial heatmaps across DINO and CLIP vision encoders, while also generating meaningful text attribution scores in CLIP’s text encoder when processing full image captions.

Table 5.4 presents quantitative segmentation results comparing SPARC’s concept-aligned approaches against established baselines.

Table 5.4: Weakly supervised segmentation results on MS COCO comparing cross-modal similarity approaches. All methods use [Chefer et al. \(2021b\)](#) relevance map generation. SPARC methods compute similarities in the concept-aligned sparse latent space, while baselines use standard feature spaces.

Method	TopK	AP	AP (M)	AP (L)	AR	AR (M)	AR (L)	mIoU
<i>Baselines</i>								
DETR	-	0.467	0.526	0.539	0.644	0.762	0.682	0.305
CLIP Similarity	-	0.248	0.370	0.282	0.420	0.422	0.625	0.157
<i>SPARC DINO Variants</i>								
Cross-Modal-Sim	Local	0.194	0.221	0.242	0.343	0.265	0.546	0.129
	Global	0.222	0.253	0.274	0.373	0.297	0.581	0.143
Concept-Latent-Sum	Local	0.212	0.243	0.251	0.349	0.283	0.518	0.136
	Global	0.222	0.244	0.264	0.352	0.284	0.516	0.137
<i>SPARC CLIP Variants</i>								
Cross-Modal-Sim	Local	0.176	0.242	0.204	0.304	0.266	0.472	0.102
	Global	0.223	0.283	0.262	0.369	0.319	0.569	0.138
Concept-Latent-Sum	Local	0.203	0.238	0.234	0.325	0.258	0.474	0.123
	Global	0.222	0.276	0.254	0.347	0.310	0.493	0.131

The results demonstrate two key insights about SPARC’s practical utility. First, Global TopK consistently outperforms Local TopK across both backbones (DINO: 0.143 vs 0.129 mIoU; CLIP: 0.138 vs 0.102 mIoU), confirming that shared activation patterns produce more coherent cross-modal representations. Second, SPARC DINO Global achieves 0.143 mIoU compared to the CLIP baseline’s 0.157 mIoU, demonstrating that text-based spatial localization through a vision-only backbone can approach the performance of natively cross-modal similarity computation when operating in SPARC’s aligned latent space.

Figure 5.4 demonstrates this using z_{279}^s (where $\mathcal{J} = \{279\}$) as the scalar target. The saliency maps show this same latent dimension responding to cat-related features across all streams: in the image (concentrated regions around the cat) and in the text (highest relevance for ”cat” token). We compare against CLIP similarity baseline, but use simplified prompts (e.g., ”a cat”) for CLIP since its cross-modal similarity naturally responds to all concepts mentioned in complex captions, while SPARC’s concept-specific latent focuses solely on the target concept even when processing full captions.

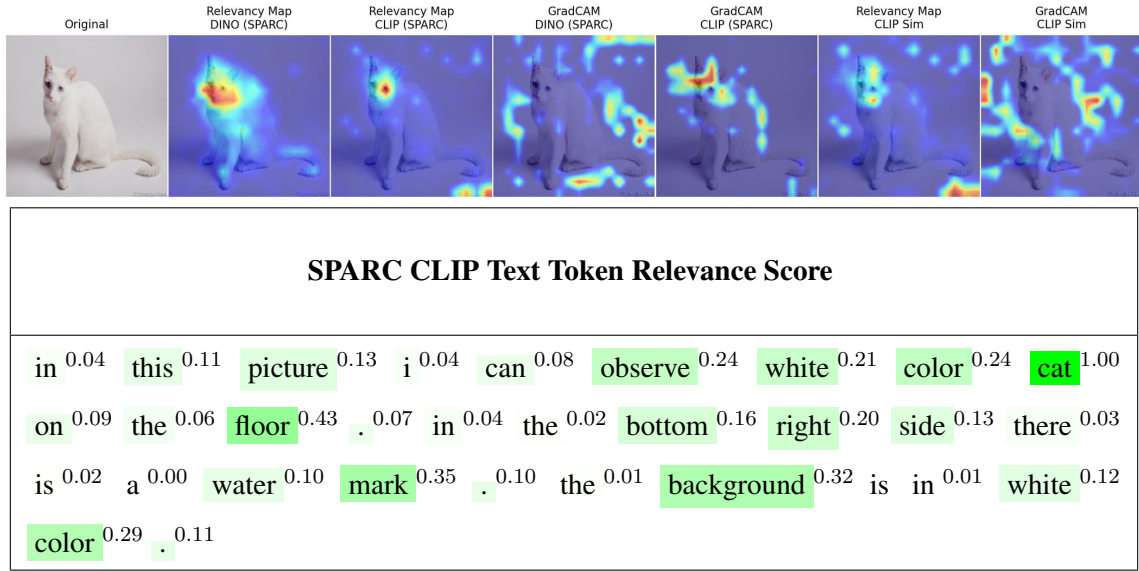


Figure 5.4: Individual latent attribution using SPARC dimension 279 vs. CLIP similarity baseline. (Above) Saliency maps show the same latent responding to cat-related features across image and text modalities. (Below) Text token relevance scores using SPARC and CLIP text.

Figure 5.5 demonstrates this using $\mathbf{z}^s \cdot \mathbf{z}^t$ computed in SPARC’s aligned latent space as scalar targets. Here, we use $\mathbf{z}^{\text{dino}} \cdot \mathbf{z}^{\text{clip.txt}}$ for DINO visualizations and $\mathbf{z}^{\text{clip.img}} \cdot \mathbf{z}^{\text{clip.txt}}$ for SPARC’s CLIP implementation, compared against the baseline CLIP similarity $\mathbf{x}^{\text{clip.img}} \cdot \mathbf{x}^{\text{clip.txt}}$. This enables text-guided spatial attention in vision-only models.

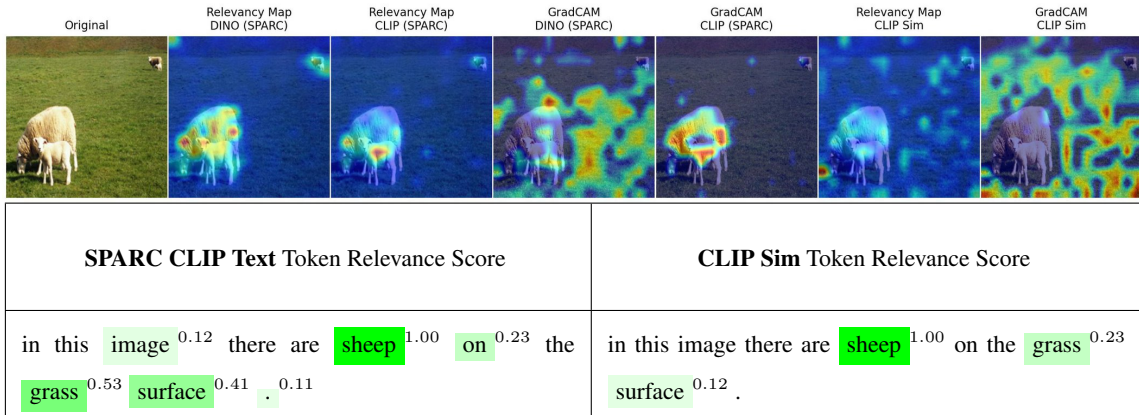


Figure 5.5: Cross-modal similarity attribution comparing SPARC’s aligned latent space against CLIP similarity baseline. Both methods process the same image and caption, showing different attribution patterns enabled by concept-aligned representations.

These examples demonstrate that SPARC’s aligned representations enable systematic concept analysis that would require separate, incompatible analyses for each model using traditional approaches. The same latent dimensions consistently represent identical concepts across different architectures and modalities, enabling direct interpretation and comparison.

5.4.5 Downstream Applications: Retrieval Performance

Table 5.5 reports R@1 scores for latent alignment across datasets and training configurations, demonstrating SPARC’s effectiveness for cross-modal retrieval tasks.

Table 5.5: Latent alignment R@1 scores across datasets and training regimes. CI = CLIP_IMG, CT = CLIP_TXT, D = DINO.

TopK	λ	CI→CT	CI→D	CT→CI	CT→D	D→CI	D→CT
<i>Open Images</i>							
Global	1	0.034	0.352	0.031	0.018	0.284	0.021
	0	0.010	0.078	0.007	0.002	0.173	0.010
Local	1	0.011	0.366	0.010	0.008	0.302	0.009
	0	0.000	0.000	0.000	0.000	0.000	0.000
<i>MS-COCO</i>							
Global	1	0.421	0.762	0.386	0.325	0.716	0.347
	0	0.181	0.259	0.204	0.102	0.369	0.188
Local	1	0.365	0.694	0.329	0.261	0.714	0.301
	0	0.000	0.000	0.000	0.000	0.000	0.000

The results reveal substantial improvements from Global TopK and cross-reconstruction loss. On MS-COCO, SPARC (Global TopK, $\lambda = 1$) achieves strong cross-modal retrieval performance: 42.1% R@1 for image-to-text and 38.6% R@1 for text-to-image retrieval. Cross-model vision retrieval also shows excellent performance, with 76.2% R@1 for CLIP-image to DINO retrieval.

The dramatic difference between configurations with and without cross-loss ($\lambda = 0$ showing 0.000 performance for Local TopK) underscores the importance of structured interaction during training for achieving meaningful cross-stream alignment.

The following parts show some of the retrieval results.

Image → Caption Retrieval

We evaluate cross-modal alignment through image-to-caption retrieval using SPARC latent representations. Tables 5.6 and 5.7 show retrieval results comparing Global vs Local TopK training configurations.

The four model configurations represent:

- **Global DINO:** Query image’s DINO features → Reference database of CLIP-text features (both from Global SPARC)
- **Local DINO:** Query image’s DINO features → Reference database of CLIP-text features (both from Local SPARC)
- **Global CLIP:** Query image’s CLIP-image features → Reference database of CLIP-text features (both from Global SPARC)
- **Local CLIP:** Query image’s CLIP-image features → Reference database of CLIP-text features (both from Local SPARC)

Each model shows top-5 retrieved captions ranked by cosine similarity in the SPARC latent space.

Model	Rank	Caption
Global DINO	1	A kite being flown in the middle of a beach.
	2	People flying kites on a sandy beach while a bucket sits in the sand.
	3	Kites being used by people on a beach.
	4	A group of people flying kites at the beach
	5	Two people on a beach flying a kite in the air.
Local DINO	1	A kite being flown in the middle of a beach.
	2	People flying kites on a sandy beach while a bucket sits in the sand.
	3	A person standing on top of a beach flying a kite.
	4	Kites being used by people on a beach.
	5	A shot of the blue water with people flying a kite.
Global CLIP	1	A kite being flown in the middle of a beach.
	2	A person standing on top of a beach flying a kite.
	3	A man is flying a kite at the beach.
	4	a man is flying a kite at on the shore at the beach
	5	Kites being used by people on a beach.
Local CLIP	1	A person standing on top of a beach flying a kite.
	2	A kite being flown in the middle of a beach.
	3	People flying kites on a sandy beach while a bucket sits in the sand.
	4	A man is flying a kite at the beach.
	5	A man flying a kite on a beach with people standing around.
Original	–	A man is flying a kite at the beach.

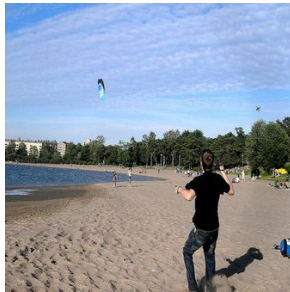


Table 5.6: The query image

is used to retrieve the captions. Green color

is for the original caption from the dataset.

5.4.6 Caption → Image Retrieval

We evaluate cross-modal alignment through caption-to-image retrieval using SPARC latent representations. Figure 5.6 shows retrieval results comparing Global vs Local TopK training configurations.

The four model configurations represent:

- **Global CLIP:** Query caption’s CLIP-text features → Reference database of CLIP-image features (both from Global SPARC)
- **Local CLIP:** Query caption’s CLIP-text features → Reference database of CLIP-image features (both from Local SPARC)
- **Global DINO:** Query caption’s CLIP-text features → Reference database of DINO features (both from Global SPARC)
- **Local DINO:** Query caption’s CLIP-text features → Reference database of DINO features (both from Local SPARC)

Each model shows top-10 retrieved images ranked by cosine similarity in the SPARC latent space.



Figure 5.6: Images retrieved for captions are (1) "In this picture we can see some food products in the glass jars.", (2) "In this image might be taken in the airplane. In this image we can see the speedometers, knobs and some digital screens." Captions are from Open Images test dataset and images are retrieved from the same split. Green boxes indicate when the corresponding image for a caption is successfully retrieved. The second caption shows such a match (Global TopK with CLIP, 3rd rank).

5.4.7 Image \rightarrow Image Retrieval

We evaluate cross-model alignment by retrieving images across DINO and CLIP-image encoders using SPARC latent representations. Figures 5.7 shows results in a 4-row layout comparing Global vs Local TopK training configurations.

The four rows represent:

- **DINO Global:** Query image's DINO features \rightarrow Reference database of CLIP features (both from Global SPARC)
- **DINO Local:** Query image's DINO features \rightarrow Reference database of CLIP features (both from Local SPARC)

- **CLIP Global:** Query image’s CLIP features \rightarrow Reference database of DINO features (both from Global SPARC)
- **CLIP Local:** Query image’s CLIP features \rightarrow Reference database of DINO features (both from Local SPARC)

Each row shows the query image (left) followed by top-10 retrieved images.

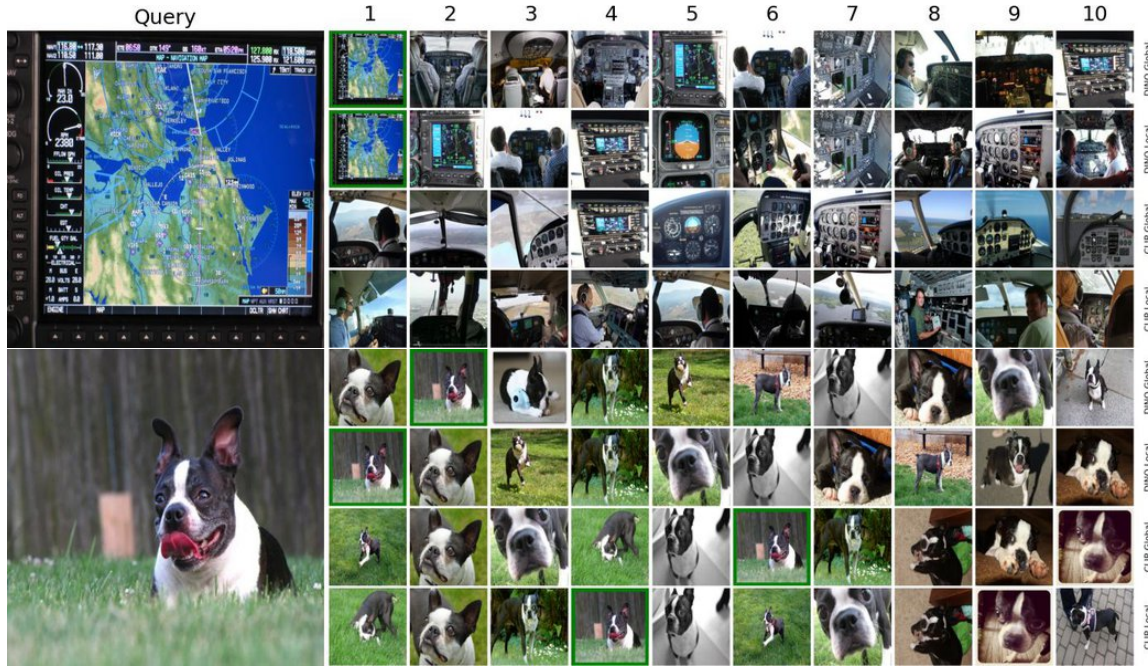


Figure 5.7: Cross-model image retrieval results. Each image shows a 4-row layout comparing query stream and reference database combinations. Query image (left) with top-10 retrieved images from reference database (right). All models trained with $\lambda = 1$ on Open Images training set. The retrieval is done on the test set of Open Images. Green border is used to show the exact match of the query image was found.

5.4.8 Ablation Analysis

Our ablation studies reveal the trade-offs inherent in SPARC’s design choices and validate the necessity of both Global TopK and cross-reconstruction components.

Figure 5.8 quantifies the cost-benefit trade-off of Global TopK constraints.

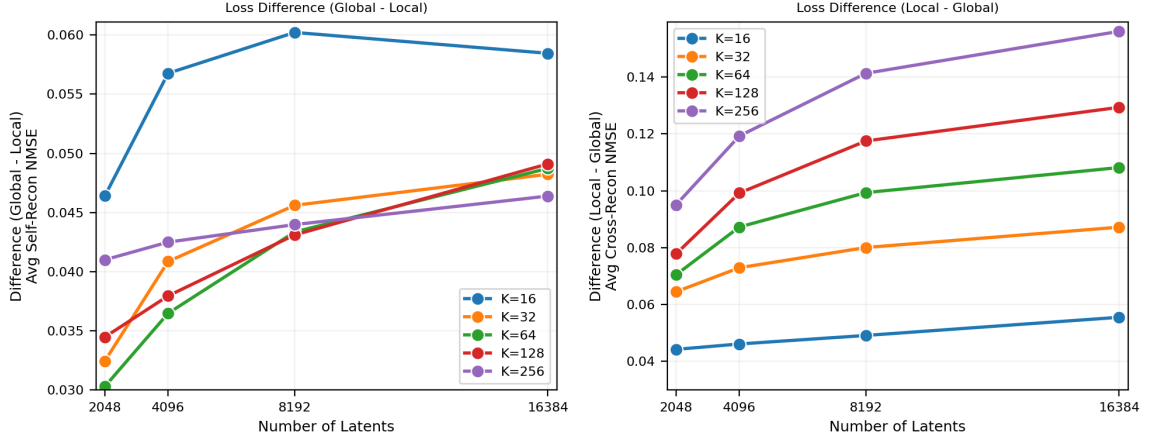


Figure 5.8: Global-vs-Local loss gap showing that Global TopK incurs self-reconstruction costs but provides larger cross-reconstruction benefits.

Global TopK incurs consistent self-reconstruction costs of 0.030-0.060 NMSE compared to Local TopK. However, these costs are more than compensated by substantial cross-reconstruction gains of 0.044-0.156 NMSE. The cross-reconstruction benefits are consistently 2-3 \times larger than the self-reconstruction penalties, indicating that the Global constraint represents a favorable trade-off for achieving concept alignment.

This quantitative analysis validates SPARC’s design philosophy: accepting modest individual model performance costs to achieve substantial gains in cross-model interpretability and alignment, directly addressing the scalability challenges that motivated this work.

Figure 5.9 shows the effect of scaling latent dimension L while keeping sparsity level k and TopK type fixed.

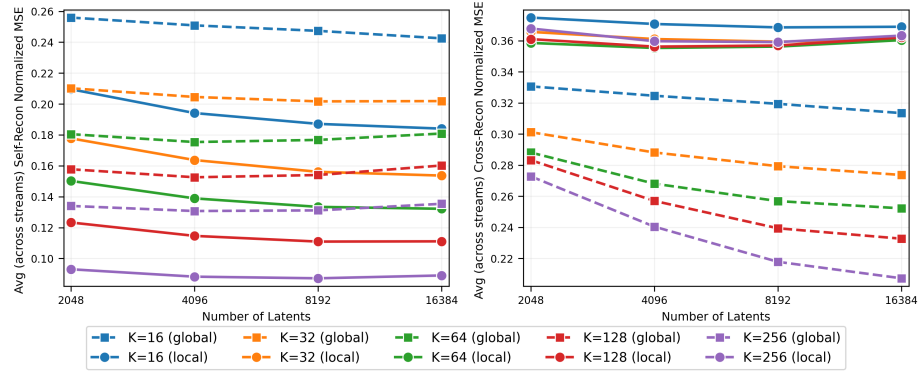


Figure 5.9: Self- and cross-reconstruction NMSE vs. number of latents, demonstrating the trade-offs between latent space capacity and reconstruction quality.

For self-reconstruction, Local TopK improves with additional latents, while Global TopK exhibits higher reconstruction loss. However, for cross-reconstruction, Global TopK demonstrates dramatic improvements when increasing the number of latents, whereas Local TopK shows much higher loss and actually degrades with additional latents. These patterns reveal that Global TopK’s structural constraint comes at the cost of self-reconstruction but delivers substantial gains in cross-reconstruction capability.

5.5 Concluding Remarks

This chapter presented SPARC as a solution to the scalability challenge identified in Chapter 4, where traditional interpretability analysis scales exponentially with model development. SPARC’s Global TopK mechanism and cross-reconstruction loss successfully transform this into a linear scaling problem by learning unified concept representations across heterogeneous models.

5.5.1 Key Achievements

Our experimental results demonstrate three critical breakthroughs. First, SPARC eliminates the incompatible latent spaces problem that plagued traditional SAE approaches, achieving 84.4% neurons active across all streams compared to 43.6% with Local TopK. Second, concept alignment improves dramatically, with Jaccard similarity reaching 0.80 versus 0.26 for traditional approaches,

more than tripling alignment quality. Third, SPARC enables previously impossible capabilities like text-guided spatial attention in vision-only models through learned cross-modal alignment.

The architectural innovation of accommodating different input dimensionalities ($d_s \neq d_t$) proves particularly valuable, enabling systematic comparison of CNN features, Transformer embeddings, and text representations within a single framework. This addresses the core limitation that prevented scalable cross-model interpretability analysis.

5.5.2 Impact on Interpretability Research

SPARC fundamentally changes how we approach cross-model interpretability. Rather than requiring separate expert analysis for each model, as demonstrated with our pathology foundation model analysis in Chapter 4, SPARC enables systematic comparison through shared latent spaces. This transformation from exponential to linear scaling directly addresses the bottleneck that limits interpretability research as model development accelerates.

5.5.3 Future Directions

Several extensions would strengthen SPARC’s practical impact. Validating the approach in specialized medical domains would directly address the pathology interpretability challenges that motivated this work. Developing methods for incremental model addition without full retraining would improve practical applicability. Finally, extending the framework to emerging architectures and modalities would ensure continued relevance as AI systems evolve.

SPARC establishes the foundation for scalable cross-model interpretability, enabling researchers to understand concept representation across diverse architectures through unified analysis rather than isolated per-model examination. This capability becomes increasingly valuable as specialized domains develop sophisticated foundation models where interpretability directly impacts practical adoption and trust.

Additional details on the SPARC implementation, along with extensive qualitative results, are provided in Appendix A.

Model	Rank	Caption
Global DINO	1	A giraffe is walking in some tall grass
	2	A giraffe standing on a grass covered field.
	3	A single giraffe looks over the green brush.
	4	there is a very tall giraffe standing in the wild
	5	a giraffe in a field with trees in the background
Local DINO	1	A giraffe standing on a grass covered field.
	2	A tall giraffe standing on top of a grass covered field.
	3	there is a very tall giraffe standing in the wild
	4	A giraffe is walking in some tall grass
	5	A giraffe standing by a pair of skinny trees.
Global CLIP	1	A giraffe stands near a tree in the wilderness.
	2	A giraffe is walking in some tall grass
	3	A giraffe standing on a grass covered field.
	4	A group of giraffes that are standing in the grass.
	5	there is a very tall giraffe standing in the wild
Local CLIP	1	A tall giraffe standing on top of a grass covered field.
	2	A giraffe standing on a grass covered field.
	3	A giraffe is walking in some tall grass
	4	A single giraffe looks over the green brush.
	5	there is a very tall giraffe standing in the wild
Original	–	A single giraffe looks over the green brush.



Table 5.7: The query image

is used to retrieve the captions.

Green color is for the original caption from the dataset.

Chapter 6

Conclusion

6.1 Summary of Contributions

This thesis presented a systematic investigation that began with efficient representation learning in medical domain and evolved into fundamental advances in cross-model interpretability. Our work makes three primary contributions that span from domain-specific medical applications to general-purpose representation engineering for vision and multimodal models.

First, we demonstrated that Vision Mamba architectures offer compelling advantages for medical applications across diverse modalities. In computational pathology, Vim4Path achieved an 8.21 AUC point improvement over Vision Transformers at comparable parameter counts, while maintaining linear computational complexity essential for processing gigapixel whole-slide images. The architecture’s sequential processing naturally aligned with pathologist workflows, as evidenced by explainability analysis showing attention to diagnostically relevant cellular features. In breast ultrasound classification, Vision Mamba models achieved statistically significant performance improvements, including 1.98% AUC and 5.0% accuracy gains on the B dataset, demonstrating effective transfer learning from natural images to medical domains.

Second, through our attempt to understand what pathology foundation models learn, we identified a fundamental scalability issue in interpretability research that extends far beyond medical domains. Our sparse autoencoder analysis of pathology models Phikon and Quilt revealed that each model produces isolated latent spaces with no inherent correspondence between dimensions.

This discovery exposed a broader challenge: while performance benchmarking scales linearly with model development, interpretability analysis scales exponentially, creating an unsustainable bottleneck as the field rapidly develops diverse foundation models.

Third, motivated by this interpretability issue, we developed SPARC (**S**parse Autoencoders for **A**ligned **R**epresentation of **C**oncepts) as a general solution for cross-model interpretability across vision and vision-language models. By learning unified latent spaces through Global TopK sparsity and cross-reconstruction losses, SPARC transforms interpretability analysis from an exponentially scaling manual process into a systematic, scalable approach. Our experiments on general vision models (DINO, CLIP) demonstrated that SPARC achieves 84.4% cross-stream neuron activation consistency and triples concept alignment quality compared to traditional approaches, while enabling new capabilities like cross-modal retrieval and weakly supervised segmentation.

6.2 Addressing the Core Challenges

The progression of this thesis, from medical-specific efficiency challenges to general interpretability solutions, addresses fundamental gaps in visual representation learning identified in our literature review.

Efficient Medical Imaging Representations. Vision Mamba architectures successfully balance computational efficiency with modeling capacity for medical domains. In pathology, both Vision Mamba and Vision Transformers process extracted patches from gigapixel WSIs rather than the full images directly. Vision Mamba’s key advantage lies in its memory efficiency, particularly during inference, where linear complexity significantly reduces memory consumption compared to Vision Transformers’ quadratic memory requirements. This memory efficiency proves crucial when processing thousands of patches per slide in clinical settings where computational resources are limited. The sequential processing naturally captures spatial relationships between neighboring tissue regions within each patch, leading to better feature representations for downstream multi-instance learning aggregation. In ultrasound, the memory-efficient architectures enable deployment on resource-constrained devices essential for point-of-care applications. These efficiency gains

came with performance improvements rather than trade-offs, suggesting that architectural alignment with domain-specific processing patterns can simultaneously optimize multiple objectives.

Scalable Interpretability. Our journey from analyzing medical models to developing SPARC revealed that interpretability challenges transcend domain boundaries. The scalability problem we discovered while studying pathology foundation models reflects a universal problem: traditional approaches requiring separate analysis of each model create exponential scaling that prevents systematic comparison. SPARC’s unified latent spaces provide a domain-agnostic solution, enabling direct concept comparison across any vision or vision-language models without model-specific analysis.

Robust Alignment Methods. SPARC demonstrates that meaningful concept alignment can be achieved across models with fundamentally different architectures, training objectives, and even input modalities. The solution emerged not from medical-specific insights but from recognizing that all models, whether processing natural images, medical scans, or text, can benefit from unified interpretability. The combination of structural constraints (Global TopK) and semantic supervision (cross-reconstruction) produces latent spaces where identical concepts activate the same dimensions across diverse models, from self-supervised vision models like DINO to multimodal models like CLIP.

6.3 Broader Implications

6.3.1 For Medical Imaging

Our medical imaging contributions demonstrate that domain-specific architectural innovations can address practical deployment challenges. Vision Mamba’s success reveals that efficiency and performance need not be traded against each other, architectures that align with clinical workflows can excel on both dimensions. The sequential processing that reduces computational complexity also produces more interpretable features, suggesting future medical AI systems should prioritize architectural choices that reflect clinical reasoning patterns. While SPARC emerged from studying medical models, its general-purpose nature means medical AI can now benefit from interpretability advances in the broader vision community. Pathology foundation models can be analyzed alongside general vision models, revealing which concepts are domain-specific versus universal across visual

understanding tasks.

6.3.2 For Vision and Multimodal AI

The progression from medical-specific challenges to general solutions illustrates how domain-focused research can yield broadly applicable insights. SPARC fundamentally changes how we approach foundation model comparison across any modalities, not just vision and vision-language. The framework’s only requirement is that inputs be paired or grouped (whether pairs, triplets, or larger sets), making it applicable to any combination of modalities: vision-audio, text-sensor data, medical imaging-genomics, or any other paired representations.

This capability becomes increasingly valuable as AI systems integrate diverse modalities. The same framework that emerged from analyzing pathology models can align concepts across arbitrary model types, whether comparing different audio encoders processing the same sound, aligning video and motion sensor representations, or understanding how various scientific models represent the same phenomena. The universality of the solution validates that interpretability challenges are fundamental to representation learning rather than specific to particular modalities or domains.

6.3.3 For Interpretability Research

Our work demonstrates that interpretability methods must evolve beyond single-model analysis to remain relevant as AI development accelerates. The journey from attempting to understand two pathology models to developing a general framework for arbitrary model comparison illustrates how practical challenges can drive fundamental methodological advances. SPARC’s approach of learning unified representations provides a template for scalable interpretability that becomes more valuable as model diversity increases. This shift from isolated to unified analysis enables new research directions in mechanistic interpretability, researchers can now study how different architectural choices and training objectives affect concept emergence across the entire landscape of vision and multimodal models, not just within narrow domains.

6.4 Limitations and Open Questions

While our contributions span from efficient medical imaging to general interpretability frameworks, several limitations merit acknowledgment.

Domain Coverage in Medical Applications. Our medical imaging experiments focused on specific applications, lymph node metastasis in pathology and breast cancer in ultrasound. While Vision Mamba showed promising results, validation across diverse pathology types and imaging modalities remains necessary. The architectural advantages we demonstrated may vary across different medical imaging characteristics.

Computational Constraints. Resource limitations affected both our medical experiments and SPARC development. Vision Mamba models might achieve even higher performance with extended training, while SPARC’s evaluation was limited to three model streams due to computational costs. Scaling to dozens of simultaneous models, the ultimate goal for comprehensive interpretability, requires further optimization.

From Medical Insights to General Methods. While SPARC successfully addresses the interpretability challenge discovered through medical model analysis, we did not complete the circle by applying SPARC back to the pathology models that motivated its development. This gap between problem identification and solution application represents an important direction for future work.

Inherent Trade-offs in Alignment. SPARC’s concept alignment comes at a modest cost to individual model reconstruction quality. This trade-off between unified understanding and individual model fidelity represents a fundamental tension. While our results suggest the benefits substantially outweigh costs for interpretability applications, use cases requiring maximum individual model performance might prefer traditional approaches.

6.5 Future Research Directions

6.5.1 Short-term Extensions

Several immediate extensions would strengthen both the medical and general interpretability contributions:

Closing the Loop on Medical Interpretability. Applying SPARC to analyze the pathology foundation models (Phikon, Quilt, UNI, Virchow) that motivated its development would validate the framework’s utility for domain-specific understanding. This would demonstrate how general interpretability solutions can enhance specialized applications.

Scaling SPARC to More Models. Extending beyond three model streams to dozens of vision and vision-language models would realize SPARC’s full potential. This includes incorporating diverse architectures (CNNs, ViTs, hybrid models), training paradigms (supervised, self-supervised, multimodal), and model scales.

Clinical Resolution Medical Imaging. Advancing Vision Mamba to clinical standards (20× magnification in pathology, real-time processing for ultrasound) while maintaining efficiency advantages would enable practical deployment. Combining these efficient architectures with SPARC’s interpretability could yield trustworthy medical AI systems.

Interactive Interpretation Interfaces. Developing tools that leverage SPARC’s unified latent spaces for interactive exploration would accelerate both research and practical applications. Users could query concept relationships across models, visualize activation patterns, and understand model disagreements through unified representations.

6.5.2 Long-term Vision

Looking beyond immediate extensions, our work points toward these possibilities:

Universal Interpretability Infrastructure. As vision and multimodal models proliferate across domains, SPARC’s approach could evolve into standard infrastructure for model understanding. Just as benchmarks enable performance comparison, unified interpretability frameworks could enable systematic concept comparison across all models.

Compositional Understanding Across Modalities. Extending unified representations beyond vision and language to include audio, sensor data, and domain-specific modalities would enable holistic understanding of multimodal AI systems. Concepts could be traced across arbitrary combinations of input modalities and model architectures.

Federated Concept Learning. SPARC’s framework could extend to federated settings where organizations collaborate to learn shared concept representations without sharing models or data.

This would enable large-scale interpretability while preserving intellectual property and privacy, critical for both commercial and sensitive applications.

Appendix A

SPARC Appendix

A.1 Experimental Details

A.1.1 Datasets

We use the dense-annotated subset of Open Images V7, containing 1.9M images (1.7M train, 41k val, 125k test), for all training and evaluation of SPARC. This subset includes bounding boxes, segmentation masks, and image-level labels. MS-COCO 2017 (118k train, 5k val) is used only for downstream segmentation and retrieval experiments.

A.1.2 Hyperparameters and Training Configuration

We provide complete hyperparameter settings for all SPARC experiments to ensure reproducibility.

Model Architecture. All experiments use a latent dimension of $L = 8,192$ with sparsity level $k = 64$. Input feature dimensions vary by dataset: for Open Images, DINOv2-ViT-L/14 features have dimension 1024 and CLIP-ViT-L-14 features have dimension 768; for MS-COCO, DINOv2-ViT-B/14 features have dimension 768 and CLIP-ViT-B/16 features have dimension 512.

Training Hyperparameters. We train all models for 50 epochs using a batch size of 256. The optimizer is Adam with learning rate $\eta = 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. All experiments use a fixed random seed of 42 for reproducibility. Training data comprises 80% of the

available samples (`train_ratio = 0.8`), with the remaining 20% reserved for validation.

Loss Function Configuration. The total loss combines self-reconstruction, cross-reconstruction, and auxiliary reconstruction terms. For cross-loss experiments, we set the cross-reconstruction coefficient $\lambda = 1.0$, while no-cross experiments use $\lambda = 0.0$.

We implement an auxiliary loss (AuxK) identical to [Gao et al. \(2025\)](#), which targets dead neurons to prevent dormant latent dimensions. Given the main reconstruction error $\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}$, the auxiliary loss reconstructs this residual using the top- k' dead latents: $\mathcal{L}_{\text{aux}} = \|\mathbf{e} - \hat{\mathbf{e}}\|_2^2$, where $\hat{\mathbf{e}} = \mathbf{W}_D^s \mathbf{z}_{\text{aux}}$. The key difference between Local and Global TopK variants is the source of the residual: Local TopK computes residuals from stream-specific reconstructions, while Global TopK computes residuals from reconstructions using globally shared indices. We use auxiliary sparsity level $k' = 64$ and coefficient $\gamma = 0.03125$ ($1/32$), with neurons considered dead after 1000 inactive steps.

Weight Initialization. Encoder and decoder weights are initialized with tied weights: $\mathbf{W}_D^s = (\mathbf{W}_E^s)^T$. Decoder weights are unit-normalized column-wise and maintained through gradient adjustments that project out components parallel to unit vectors. All bias parameters are initialized to zero. Dead neurons are reinitialized using Gaussian noise with a standard deviation of 0.01.

Dead Neuron Management. Latent dimensions with activation frequency below 10^{-3} for more than 1000 consecutive training steps are considered dead and reinitialized. This prevents the emergence of inactive dimensions during training while the auxiliary loss encourages their reactivation.

Data Loading Optimization. To address HDF5 loading efficiency, we implement a custom `ContiguousRandomBatchSampler` that shuffles at the batch level rather than sample level, reducing training time from hours to minutes per epoch while maintaining training dynamics. `DataLoaders` use 4 workers with memory pinning enabled.

Feature Model Specifications. For Open Images experiments, we extract features using DINOv2-ViT-L/14 (with registers variant) and CLIP-ViT-L-14 trained on DataComp-1B. For MS-COCO experiments, we use DINOv2-ViT-B/14 (with registers variant) and CLIP-ViT-B/16 trained on DataComp-XL dataset (`datacomp_xl_s13b_b90k`).

Experimental Configurations. We evaluate four training configurations across two datasets, resulting in eight total experimental conditions: Global/Local TopK activation \times Cross-loss/No

Cross-loss \times Open Images/MS-COCO datasets. The Global TopK variant aggregates logits across streams before index selection, while Local TopK applies independent TopK selection per stream.

A.1.3 Probe Implementation Details

We evaluate concept recoverability using 1D logistic probes following the methodology of [Gao et al. \(2025\)](#). This section details the complete experimental procedure for the Open Images binary classification evaluation reported in Section 5.4.3.

Dataset and Task Selection. We use the Open Images test set containing 112,699 samples with 601 available binary classification labels. To ensure statistical reliability, we filter tasks to retain only those with at least 50 positive examples, resulting in 432 binary classification tasks used in our evaluation.

Data Preprocessing and Balancing. For each binary classification task, we address class imbalance by randomly sampling negative examples to match the number of positive examples, creating balanced datasets. Latent activations are binarized using a threshold of zero: sparse representations \mathbf{z}^s are converted to binary indicators ($z_i^s > 0$) for probe training.

Data Splitting Strategy. We employ a stratified three-way split with 70% training, 15% validation, and 15% test samples from the balanced dataset. Stratification ensures both positive and negative classes are represented proportionally across all splits. We use task-specific random seeds (`1000 + task_id`, `2000 + task_id`, `3000 + task_id`) for reproducible splits across experiments.

Candidate Latent Selection. Since our goal is to test whether individual latent dimensions can recover semantic concepts, we focus the evaluation on the most promising candidates rather than all 8,192 latent dimensions. We first exclude dimensions that show zero activation during training, then rank remaining candidates by counting how many positive training examples activate each dimension. For each task, we select the 20 most frequently activated dimensions as candidates, reducing computational cost while focusing on latents most likely to encode the target concept.

Probe Training Configuration. For each candidate latent dimension, we train a 1D logistic regression probe with the following hyperparameters: L-BFGS solver, maximum 200 iterations, L2 regularization with default strength ($C = 1.0$), single latent activation value z_i^s as input, and binary class labels as targets.

Model Selection and Evaluation. We select the best-performing latent dimension for each task based on validation set performance. For each of the 20 candidate dimensions, we train individual probes, evaluate on the validation set using binary cross-entropy loss, select the dimension achieving the lowest validation loss, and report final performance on the held-out test set.

Evaluation Metrics. We report mean binary cross-entropy loss across all 432 tasks for each stream and training configuration. As a baseline reference, random binary classification on balanced data yields an expected loss of $-\log(0.5) \approx 0.693$.

Experimental Scope. We evaluate four SPARC training configurations (Global/Local TopK \times Cross-loss/No Cross-loss) across three feature streams (DINO, CLIP-image, CLIP-text), resulting in 12 total experimental conditions. Each condition processes all 432 binary classification tasks using the identical probe methodology described above.

A.2 Latent Dimension Visualizations

This section provides examples of latent activation patterns across training configurations (Local/Global TopK \times $\lambda=0/1$). Each figure shows the 2 \times 2 grid, with each configuration displaying top-10 activating images across three streams (DINO, CLIP-image, CLIP-text) for latent dimensions.

Figures A.1, A.2, and A.3 demonstrate concept alignment under Global TopK with $\lambda = 1$, where latent dimensions exhibit consistent activation behavior across all streams—either fully active or completely dead across all three modalities. In contrast, Local TopK with $\lambda = 1$ shows mixed activation patterns, where latents may be active in some streams while remaining inactive in others. For more results, check <https://github.com/AtlasAnalyticsLab/SPARC/blob/main/VISUALIZATIONS.md>.

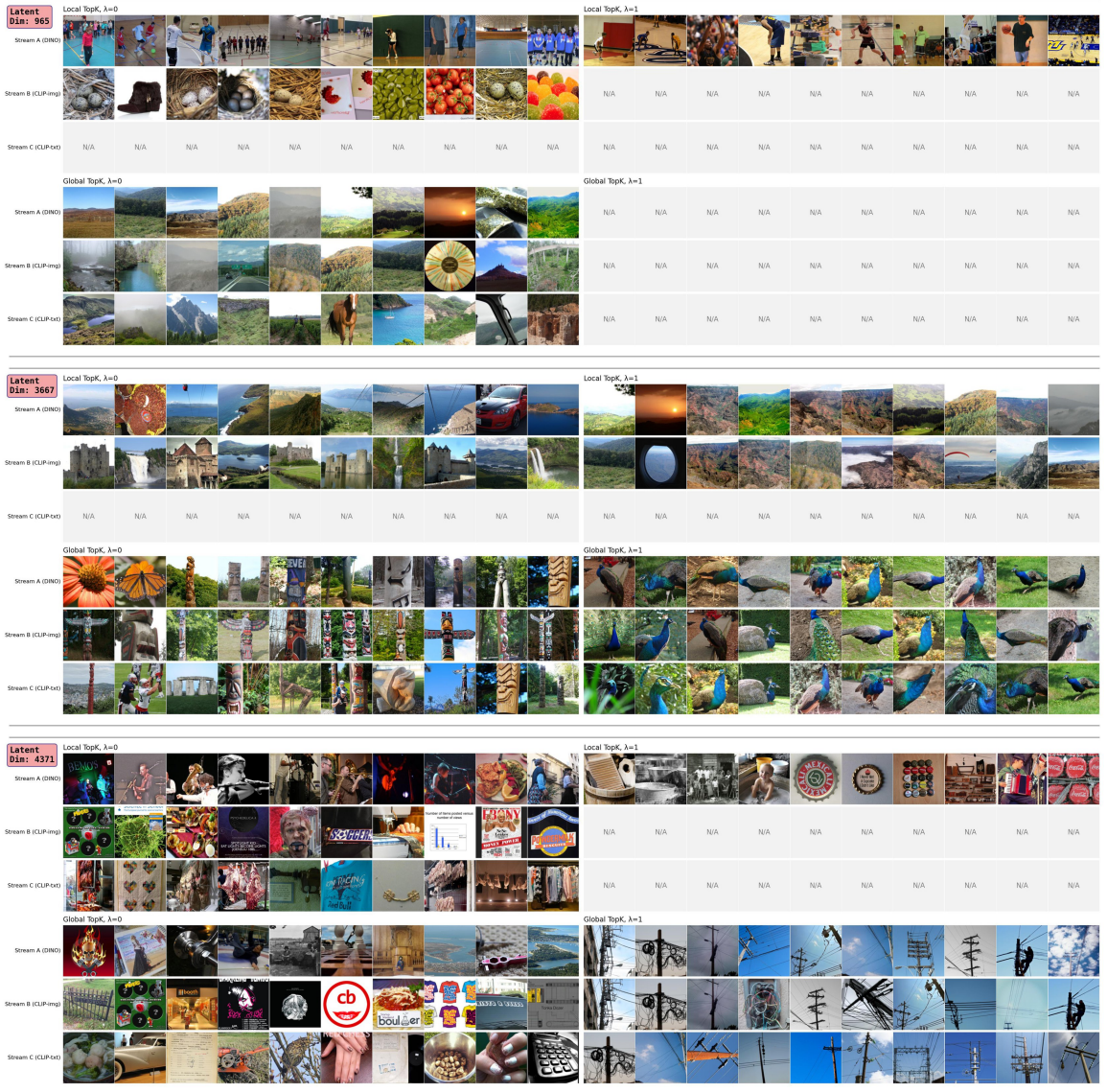


Figure A.1: Latent activation examples for dimensions 965, 3667, and 4371 showing top-10 activating images across different configurations.

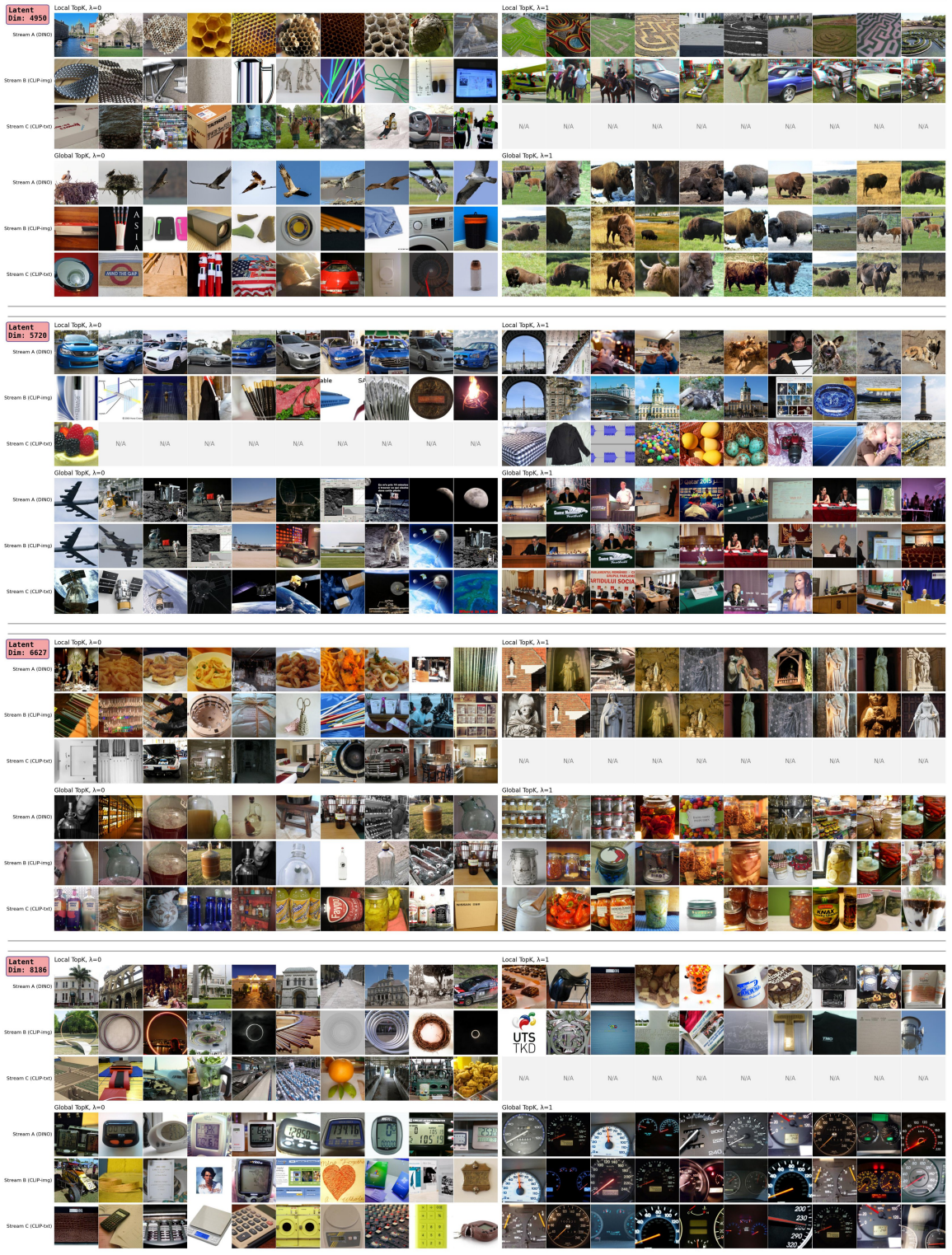


Figure A.2: Additional latent activation examples for dimensions 4950, 5720, 6627, and 8186 showing top-10 activating images across different configurations.

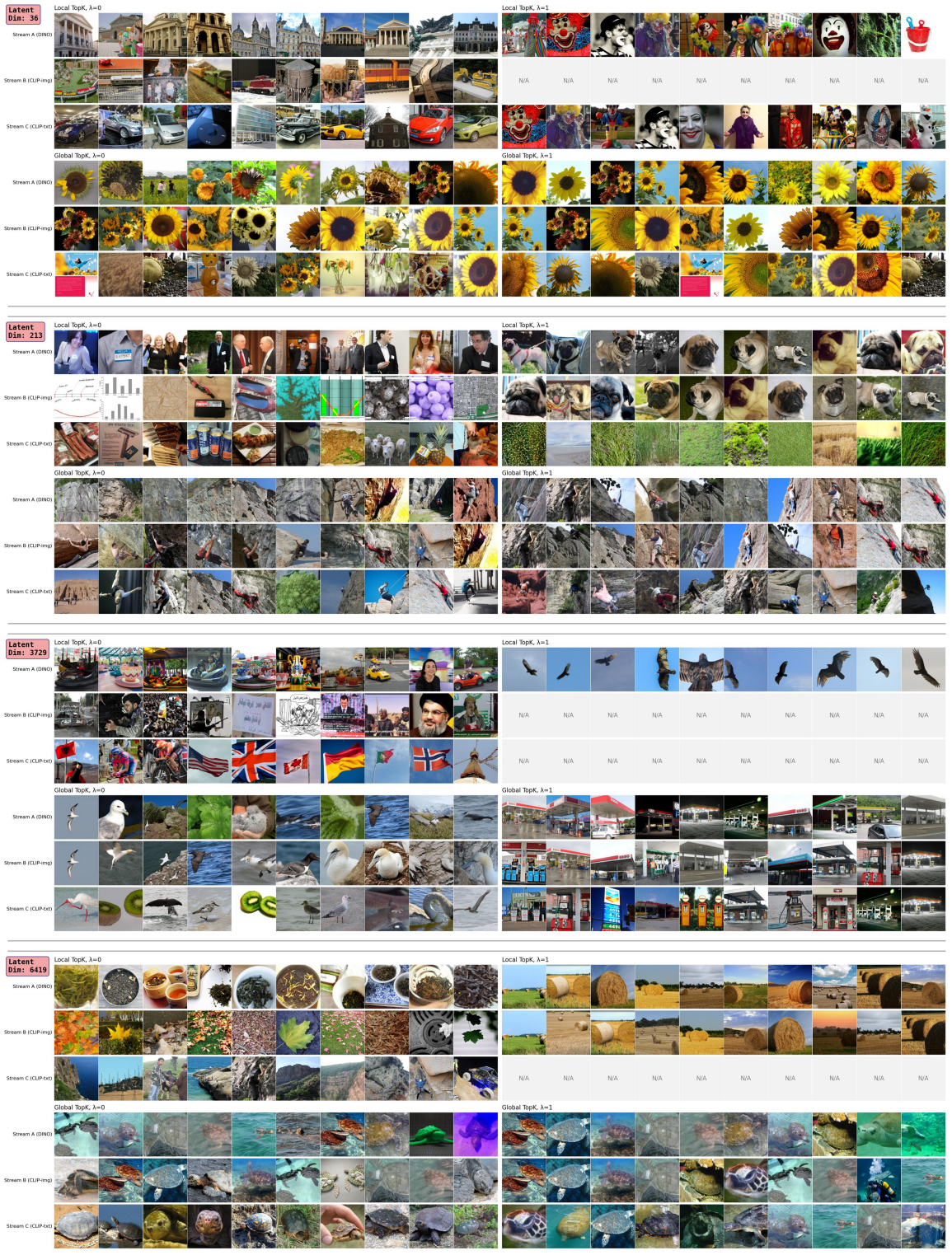


Figure A.3: Additional latent activation examples for dimensions 36, 213, 3729, and 6419 showing top-10 activating images across different configurations.

A.3 Latent Attribution based on concepts

This section demonstrates attribution analysis using SPARC’s concept-aligned latents as scalar targets for gradient-based methods. We identify concept-relevant latents by analyzing their activation patterns on labeled data, then use their combined activations as attribution targets for spatial and textual attribution.

To identify which concepts each latent represents, we examine what types of images most strongly activate each latent dimension. For each latent, we analyze its top-50 activating samples and determine the most frequent concept category among them. We measure semantic consistency by computing the purity score—the fraction of top-activating samples that belong to the dominant category. When this purity exceeds 0.3, indicating sufficient semantic consistency, we assign the corresponding concept name to that latent. Latents without clear semantic patterns remain unlabeled. This process enables identification of multiple latents representing the same concept across different streams.

For attribution, we collect all latents assigned to a target concept across streams, forming the set \mathcal{J} of concept-relevant indices. We then compute spatial attribution using both relevancy maps (Chefer, Gur, & Wolf, 2021a) and GradCAM (Selvaraju et al., 2017) with the summed activations $\sum_{j \in \mathcal{J}} z_j^s$ as scalar targets. For text attribution, we compute token relevancy scores (Chefer et al., 2021a). Each figure displays spatial attribution heatmaps alongside text token relevance scores, with scores below 0.1 omitted for clarity. For more results, check <https://github.com/AtlasAnalyticsLab/SPARC/blob/main/VISUALIZATIONS.md>.

A.3.1 Concept Specific Latents

Figures A.4, A.5, A.6, and A.7 show attribution results using concept-specific latent selections across various object categories.

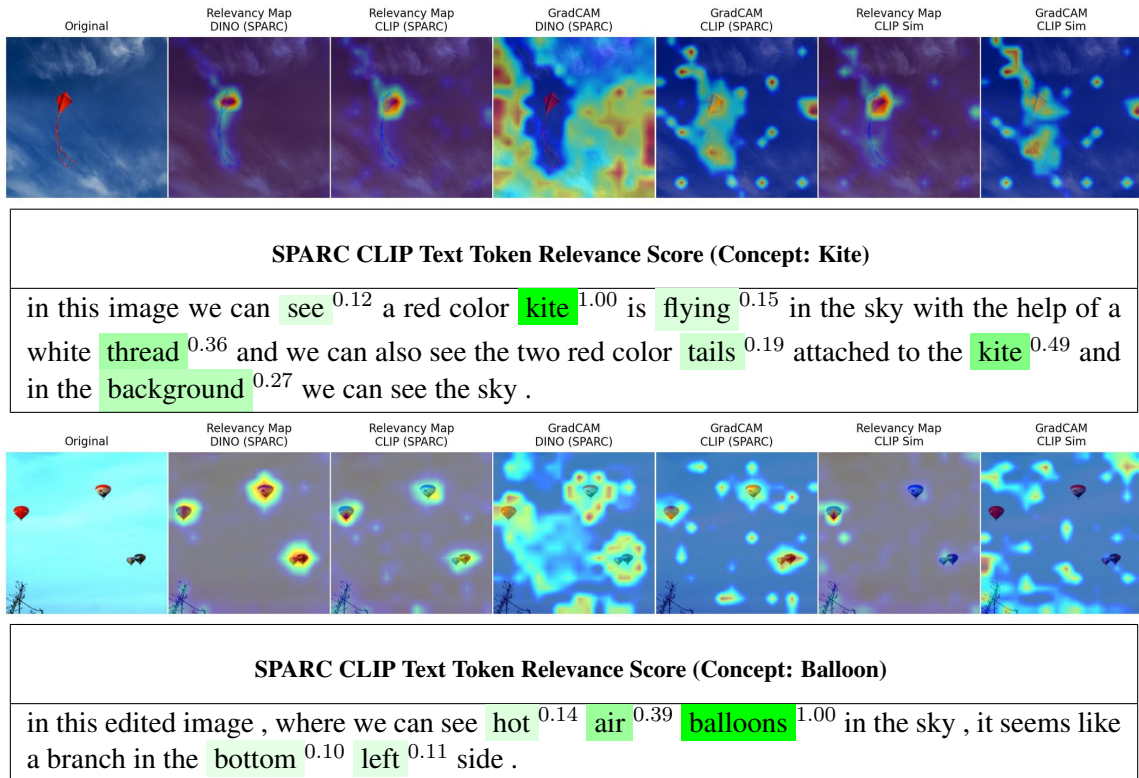


Figure A.4: SPARC CLIP text token relevance for kite and balloon concepts. CLIP similarity baseline uses concept names "a kite" and "a balloon" rather than full captions.

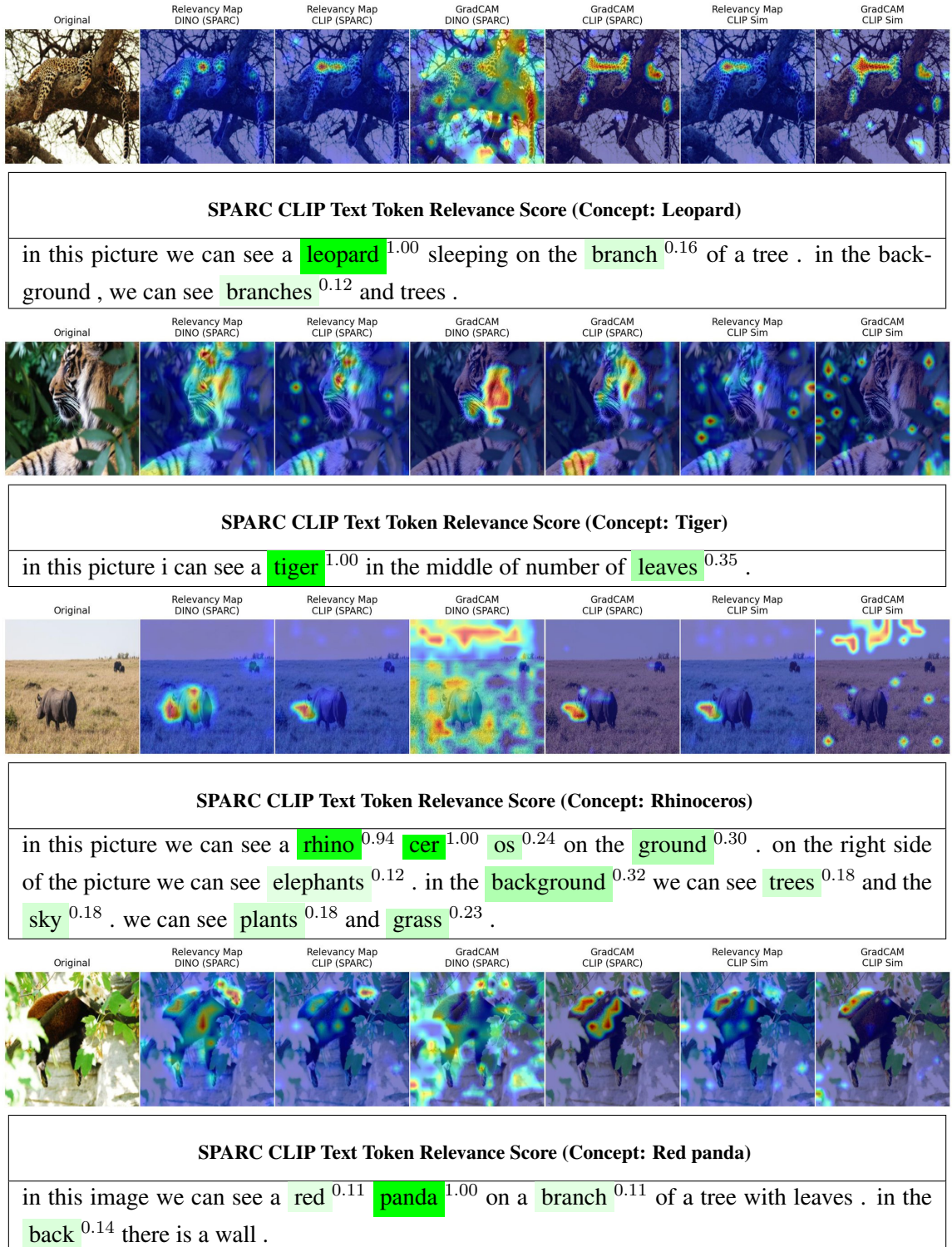
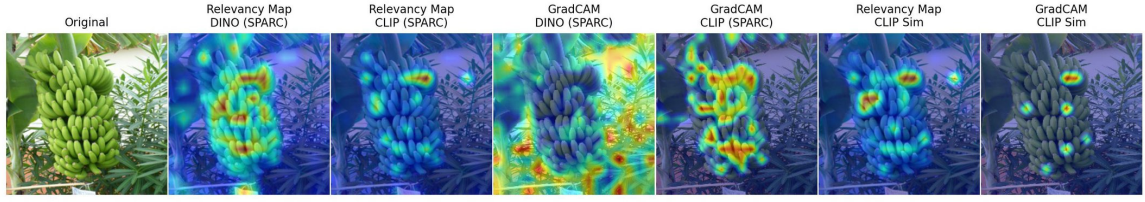
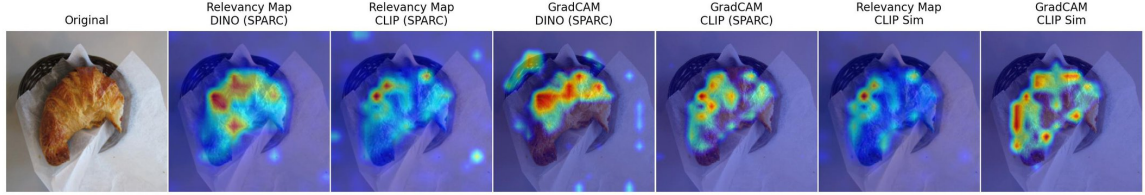


Figure A.5: SPARC CLIP text token relevance for leopard, tiger, rhinoceros, and red panda concepts. CLIP similarity baseline uses concept names "a leopard", "a tiger", "a rhinoceros", and "a red panda" rather than full captions.



SPARC CLIP Text Token Relevance Score (Concept: Banana)

in this picture we can see **bananas**^{1.00} , there are some plants and in the **background**^{0.18} of the picture there is a wall .



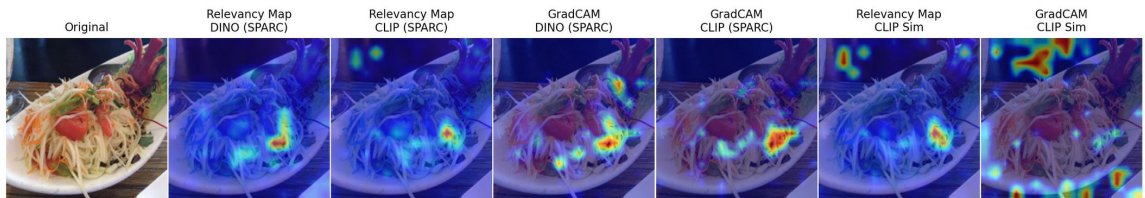
SPARC CLIP Text Token Relevance Score (Concept: Croissant)

in this **image**^{0.11} there is a **bowl**^{0.14} on a **surface**^{0.21} , in that **bowl**^{0.11} there are **tissue**^{0.13} **papers**^{0.26} and a **croissant**^{1.00} .



SPARC CLIP Text Token Relevance Score (Concept: Cake)

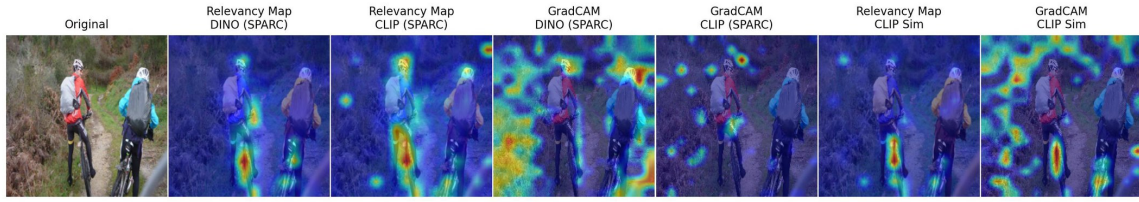
at the bottom of this image , there are some **cakes**^{1.00} arranged on a table . in the middle of this image , there is a woman in a **violet**^{0.12} colorful t - **shirt**^{0.28} having a **badge**^{0.20} , **smiling**^{0.13} and standing . in the background , there is a white color **curtain**^{0.34} , a **wall**^{0.10} and other **objects**^{0.24} .



SPARC CLIP Text Token Relevance Score (Concept: Pasta)

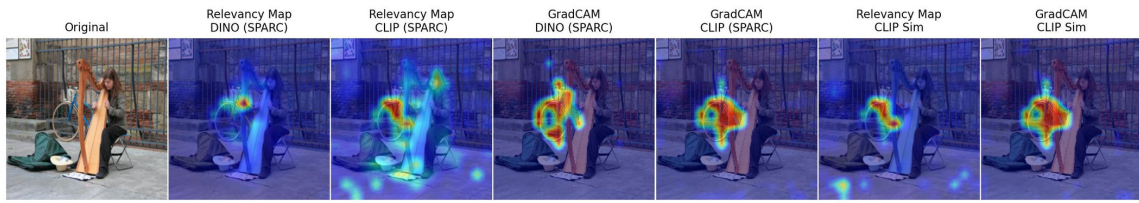
in this image we can see **noodles**^{1.00} and **vegetables**^{0.12} in **plate**^{0.12} on the table . to the right side of the image there is a glass .

Figure A.6: SPARC CLIP text token relevance for banana, croissant, cake, and pasta concepts. CLIP similarity baseline uses concept names "a banana", "a croissant", "a cake", and "pasta" rather than full captions.



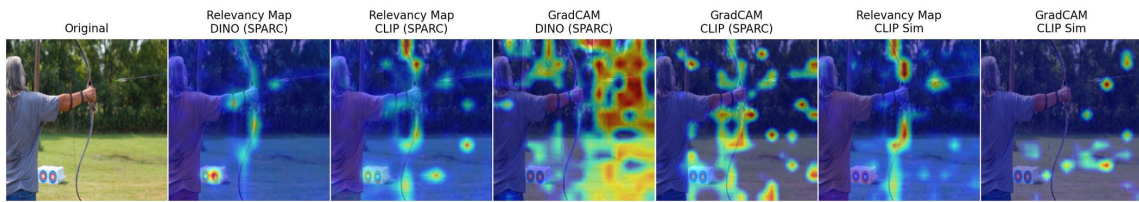
SPARC CLIP Text Token Relevance Score (Concept: Bicycle)

in this image we can see two **people**^{0.16} **riding**^{0.19} **bicycles**^{1.00} . in the background of the image there are **trees**^{0.16} , **plants**^{0.15} , a **pole**^{0.20} , **grass**^{0.10} and other **objects**^{0.17} . at the bottom of the image there is the **grass**^{0.18} and **ground**^{0.13} . on the right side **bottom**^{0.11} of the image there is an **object**^{0.13} .



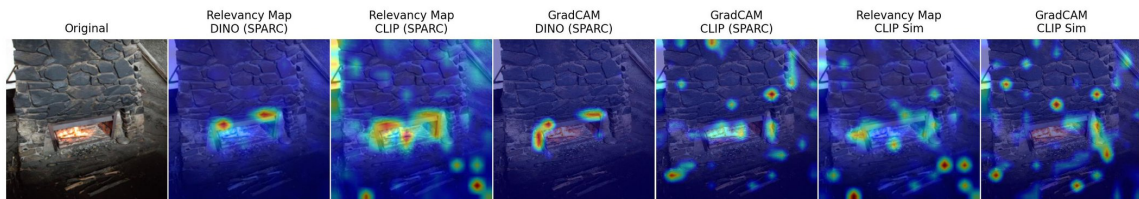
SPARC CLIP Text Token Relevance Score (Concept: Bicycle)

in this image we can see a **woman**^{0.13} holding a **musical**^{0.15} **instrument**^{0.18} . we can see a **bag**^{0.11} , **hat**^{0.21} and some **objects**^{0.12} . in the background we can see the **fence**^{0.18} , **bicycle**^{1.00} , walls and **frames**^{0.20} .^{0.12}



SPARC CLIP Text Token Relevance Score (Concept: Bow and arrow)

in this image , i can see a person standing and holding a **bow**^{1.00} and there is an **arrow**^{0.64} in the air . at the bottom of the image , i can see an **object**^{0.15} on the **grass**^{0.14} . in the **background**^{0.16} there are **trees**^{0.12} and a **pole**^{0.15} .



SPARC CLIP Text Token Relevance Score (Concept: Wood-burning stove)

in this **picture**^{0.21} we can **see**^{0.19} few **sticks**^{1.00} and few **objects**^{0.35} on the **ground**^{0.21} and in the **background**^{0.20} we can **see**^{0.15} the **fire**^{0.55} **place**^{0.76} and the **wall**^{0.32} .

Figure A.7: SPARC CLIP text token relevance for bicycle, bow and arrow, and wood-burning stove concepts. CLIP similarity baseline uses concept names "a bicycle", "bow and arrow", and "wood-burning stove" rather than full captions.

A.3.2 Same image/caption, different latents

Figure A.8 demonstrates how different concept selections produce attribution patterns for identical inputs. By changing the set \mathcal{J} while keeping same image and caption, SPARC generates different attribution patterns.

A.3.3 Using latents of concepts that are not present in the image/caption

Figure A.9 examines SPARC’s behavior when applying concept latents to samples lacking those concepts. For these samples, when using irrelevant concept latent sets \mathcal{J} , SPARC latents will be zero, meaning no gradients are produced and all attribution scores remain zero.

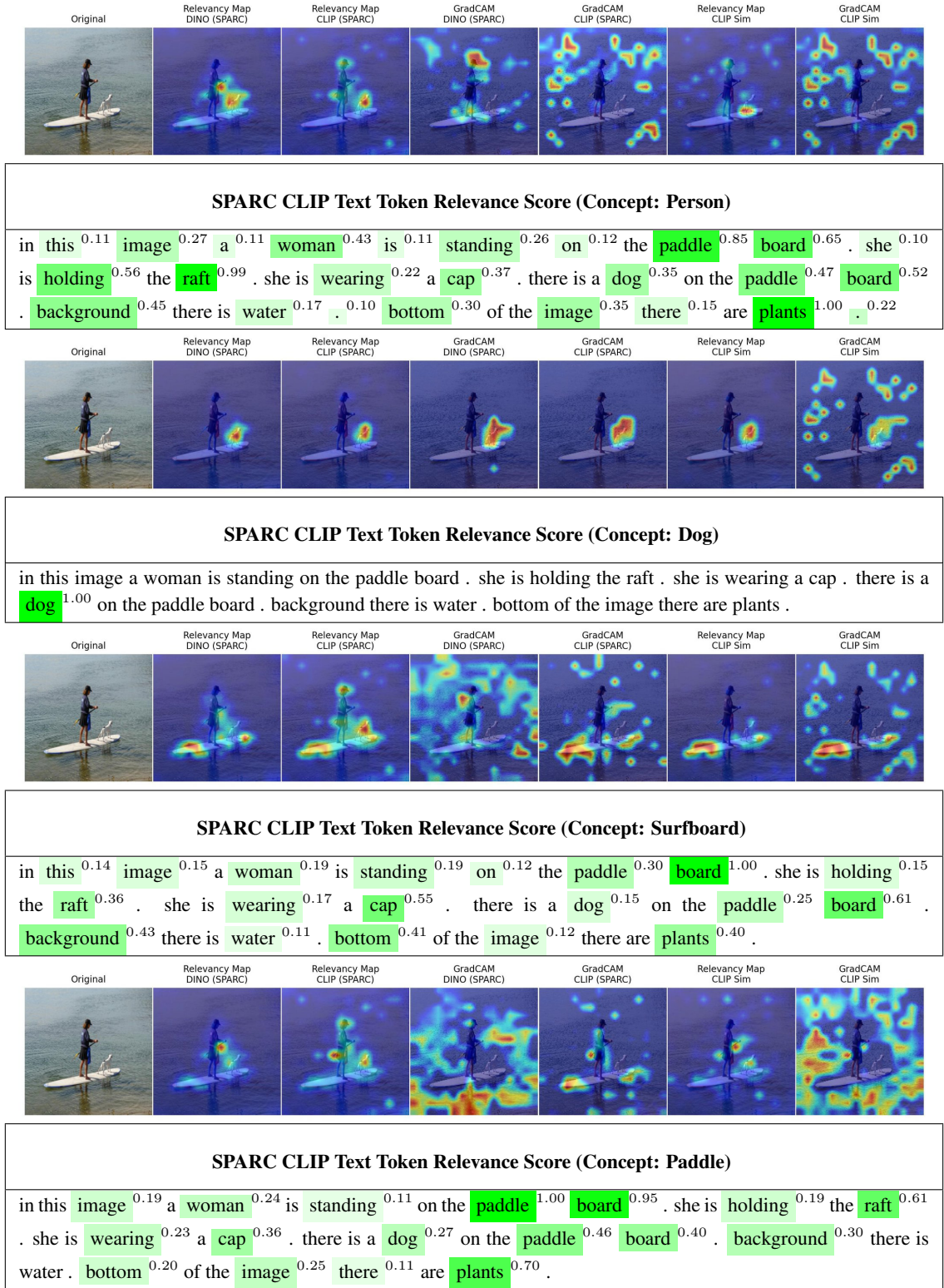


Figure A.8: SPARC CLIP text token relevance for the same image/caption using different concept-specific latent sets. Each panel shows attribution for a different target concept as indicated in the table headers.



SPARC CLIP Text Token Relevance Score (Concept: Cat)

in this image a woman is standing on the paddle board . she is holding the raft . she is wearing a cap . there is a dog on the paddle board . background there is water . bottom of the image there are plants .



SPARC CLIP Text Token Relevance Score (Concept: Apple)

at the bottom of this image , there are food items arranged on a table . in the middle of this image , there is a woman in a violet colorful t - shirt having a badge , smiling and standing . in the background , there is a white color curtain , a wall and other objects .



SPARC CLIP Text Token Relevance Score (Concept: Flag)

in this image we can see a woman holding a musical instrument . we can see a bag , hat and some objects . in the background we can see the fence , bicycle , walls and frames .



SPARC CLIP Text Token Relevance Score (Concept: Tiger)

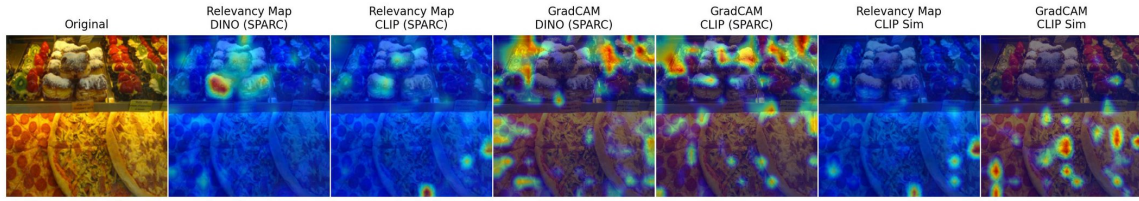
in this picture we can see bananas , there are some plants and in the background of the picture there is a wall .

Figure A.9: SPARC text token relevance for image/captions with a concept that's not present in the sample. Using irrelevant latent dimensions in SPARC causes no gradients. For text scores, all scores are 0.

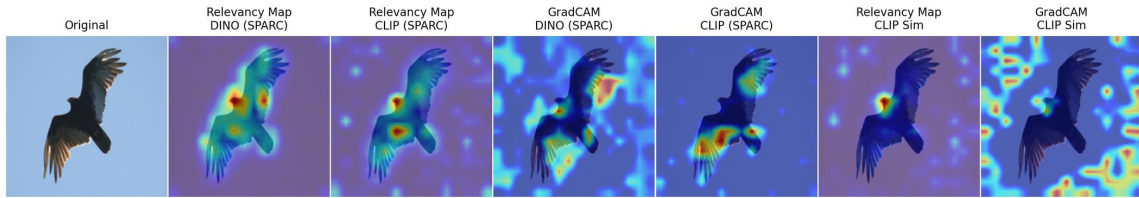
A.3.4 Limitations of concept-based latent attribution

For certain common classes such as person and car, many latents receive concept assignments—over 500 latents for "person" and nearly 300 for "car". This is mainly a limitation of assigning concepts to latents as discussed in Section [5.4.2](#).

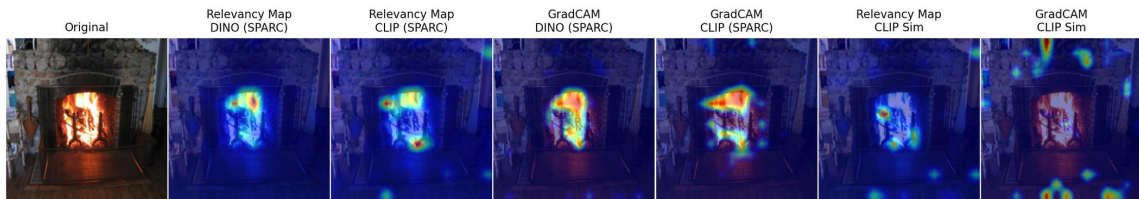
This breaks the selective attribution behavior of SPARC demonstrated in Appendix [A.3.3](#). While concept-specific latents for most classes produce no gradients when the concept is absent, common concepts with numerous assigned latents produce spurious activations even when not present. Figure [A.10](#) demonstrates this limitation, where "person" and "car" latents generate non-zero attributions on images containing neither concept.



SPARC CLIP Text Token Relevance Score (Concept: Person)									
in	0.01	this	0.02	image	0.11	there	0.03	are	0.02
				food	0.29	items	0.17	in	0.03
				the	0.03	trays	0.17		
which	0.02	are	0.01	placed	0.03	on	0.03	the	0.12
				glass	0.12	display	0.11	with	0.05
								the	0.05
								tags	1.00
.									



SPARC CLIP Text Token Relevance Score (Concept: Car)									
in	0.06	this	0.14	picture	0.31	i	0.19	can	0.34
				see	0.38	an	0.15	eagle	1.00
				in	0.08	the	0.08	sky	1.00
.	0.19								



SPARC CLIP Text Token Relevance Score (Concept: Person)									
in	0.02	the	0.02	foreground	0.18	of	0.02	the	0.01
				image	0.04	we	0.02	can	0.02
				see	0.11	the	0.01		
floor	0.10	and	0.01	metal	0.05	rods	0.27	.	0.01
				on	0.01	the	0.10	left	0.10
				side	0.02	we	0.01	can	0.01
see	0.06	a	0.33	table	0.33	on	0.01	which	0.01
				books	0.07	are	0.01	there	0.01
				, a	0.02	blue	0.02	color	0.03
box	0.08	, a	0.03	wooden	0.03	object	0.04	and	0.01
				a	1.00	poster	1.00	like	0.03
				structure	0.08	are	0.01		
there	0.01	.	0.01	on	0.01	the	0.25	right	0.25
				side	0.03	we	0.01	can	0.01
				see	0.06	a	0.06	broom	0.06
stick	0.29	and	0.07	a	0.07	wall	0.07	.	0.01
				in	0.01	the	0.03	middle	0.03
				we	0.02	can	0.01	see	0.06
				stones	0.07				
wall	0.11	on	0.01	which	0.02	a	0.01	white	0.06
				color	0.05	object	0.09	is	0.02

Figure A.10: SPARC CLIP text token relevance for image/captions with a concept that's not present in the sample. SPARC produces non-zero gradients for some of common concepts even in the absence of the concept.

A.4 Cross-Modal Similarity Attribution

This section demonstrates attribution using cross-modal similarities $\mathbf{z}^s \cdot \mathbf{z}^t$ in SPARC’s aligned latent space as scalar targets. We compute spatial attribution using both relevancy maps (Chefer et al., 2021a) and GradCAM (Selvaraju et al., 2017), while text attribution uses token relevancy scores (Chefer et al., 2021a).

A.4.1 Cross-modal heatmaps with full captions

Figures A.11, A.12, and A.13 demonstrate cross-modal attribution using full captions as text input. Each figure shows spatial heatmaps alongside text token relevance scores, comparing SPARC’s aligned latent similarities against CLIP similarity baselines across various object categories. For more results, check <https://github.com/AtlasAnalyticsLab/SPARC/blob/main/VISUALIZATIONS.md>.

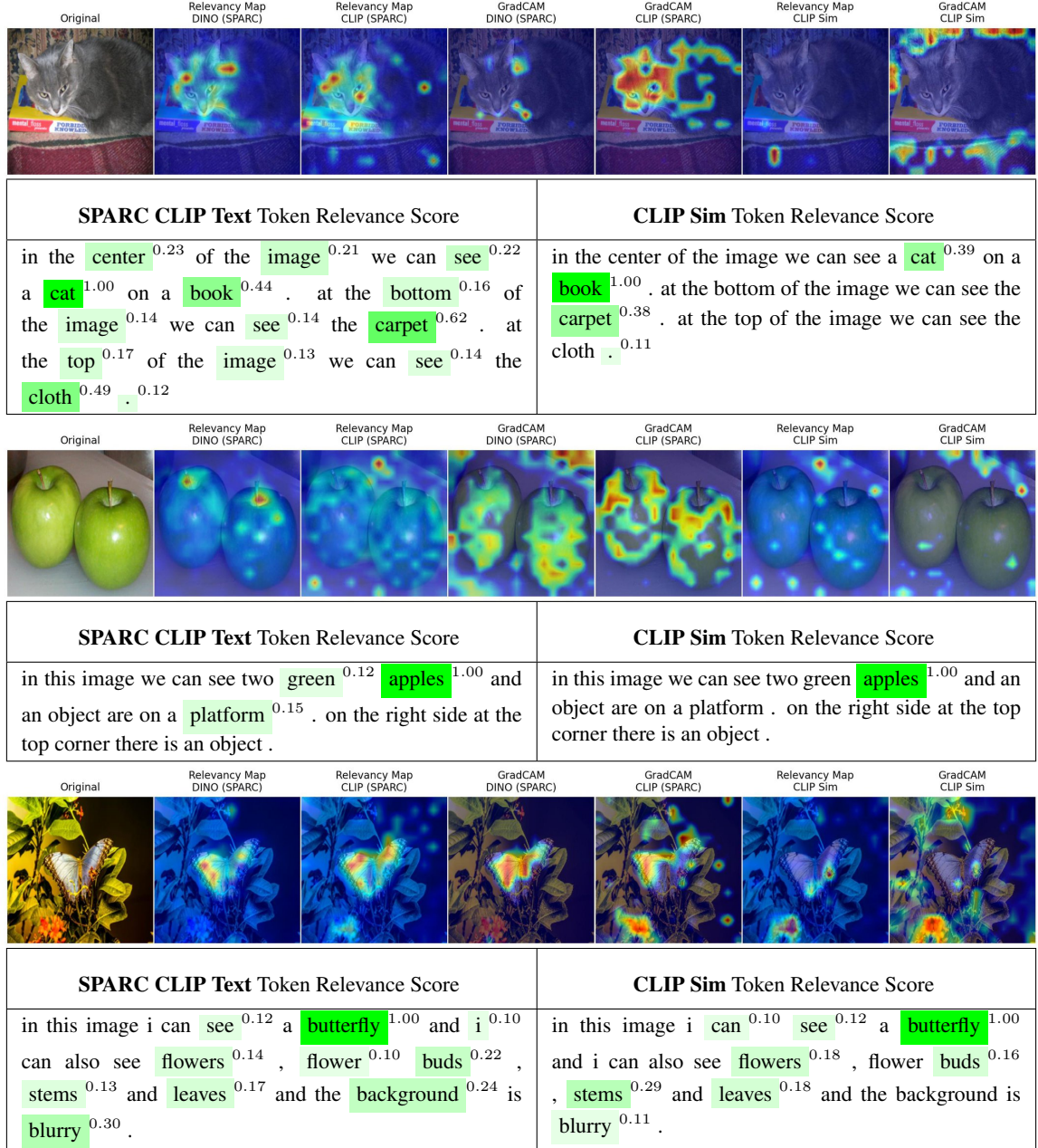


Figure A.11: Cross-modal similarity attribution for mixed concepts (cat, apple, butterfly) comparing SPARC’s aligned latent space against CLIP similarity baseline.

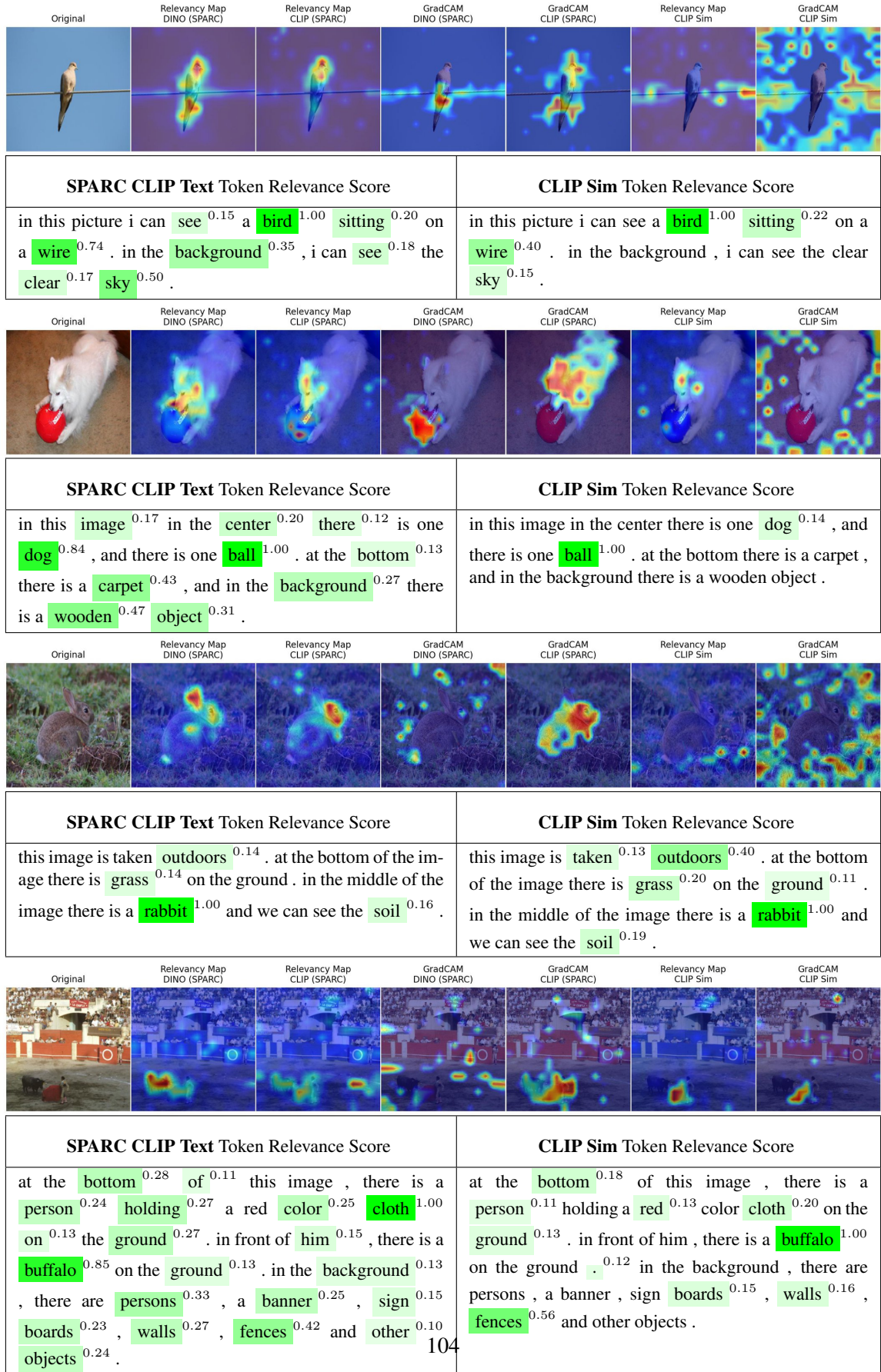


Figure A.12: Cross-modal similarity attribution for animal concepts comparing SPARC’s aligned latent space against CLIP similarity baseline.

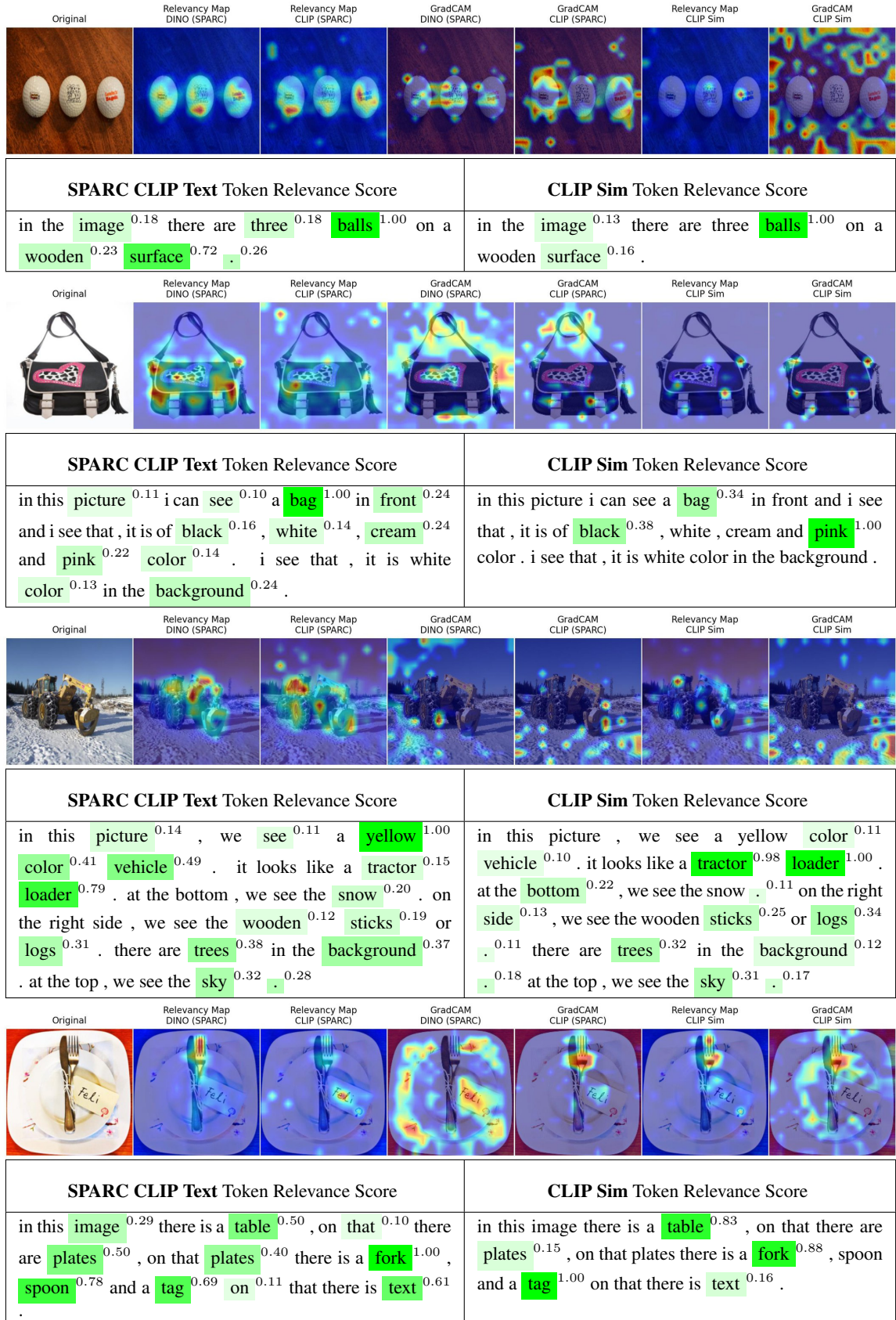


Figure A.13: Cross-modal similarity attribution for object concepts comparing SPARC's aligned latent space against CLIP similarity baseline.

A.4.2 Same image, different captions

Figure A.14 demonstrates how different text queries produce distinct spatial attribution patterns when applied to the same image. Using simple concept names ("Banana", "Apple", "Kiwi", "Cat") as text inputs, we examine how cross-modal similarities $\mathbf{z}^s \cdot \mathbf{z}^t$ generate different heatmaps based on text guidance, including cases where the queried concept is absent from the image.

For this sample image, CLIP similarity shows focused attribution on target objects when present in the image. SPARC exhibits more varied attribution patterns, localizing to different image regions with less precision than CLIP similarity. We make no claims about performance or whether these heatmaps are meaningful or relevant to what the actual encoders are looking at. We present these examples to demonstrate that SPARC's spatial attribution varies with different text inputs, indicating text-guided behavior rather than text-independent object highlighting.

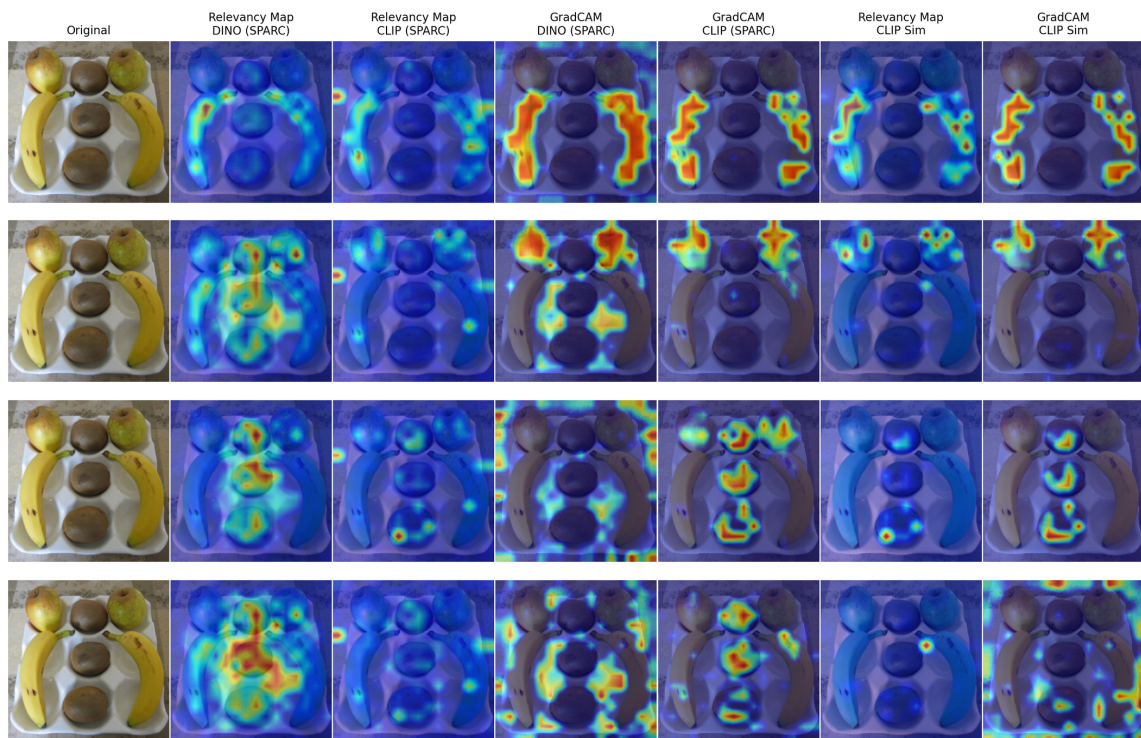


Figure A.14: Captions used are "Banana", "Apple", "Kiwi", and "Cat" (non-existent concept).

A.4.3 Cross-modal attribution limitations

Figure A.15 shows a case where SPARC’s spatial attribution produces less meaningful localization compared to Figure A.14. Using detailed spatial queries (“Cat’s Ears”, “Cat’s Eyes”, “Cat’s Nose”, “A Cat”), SPARC generates similar, poorly localized attribution patterns with minimal variation across different text inputs. Although CLIP doesn’t match the captions perfectly, it shows more responsiveness to the text input.

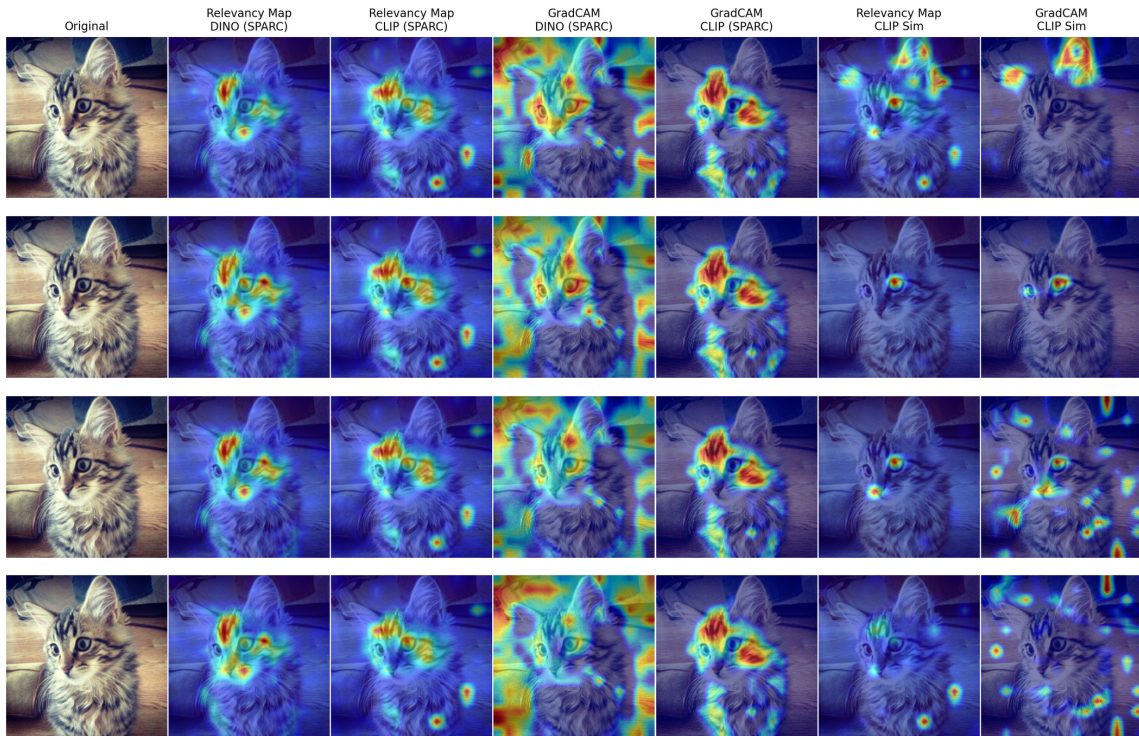


Figure A.15: Captions used are “Cat’s Ears”, “Cat’s Eyes”, “Cat’s Nose”, and “A Cat”. We find SPARC fails in the case of detailed heatmaps.

A.5 Retrieval Qualitative Results

This section presents qualitative retrieval examples using SPARC’s aligned latent representations. The following subsections show retrieval samples across in-distribution tasks using test data queries, and out-of-distribution tasks using external queries not present in the dataset. In both cases, the reference database consists of test set samples.

A.5.1 Image → Caption Retrieval (In-Distribution)

We evaluate cross-modal alignment through image-to-caption retrieval using SPARC latent representations. Tables [A.1–A.10](#) show retrieval results comparing Global vs Local TopK training configurations.

The four model configurations represent:

- **Global DINO:** Query image’s DINO features → Reference database of CLIP-text features (both from Global SPARC)
- **Local DINO:** Query image’s DINO features → Reference database of CLIP-text features (both from Local SPARC)
- **Global CLIP:** Query image’s CLIP-image features → Reference database of CLIP-text features (both from Global SPARC)
- **Local CLIP:** Query image’s CLIP-image features → Reference database of CLIP-text features (both from Local SPARC)

Each model shows top-5 (or top-4) retrieved captions ranked by cosine similarity in the SPARC latent space.

Open Images dataset’s Image → Caption Retrieval

Tables [A.1](#), [A.2](#), [A.3](#), [A.4](#), and [A.5](#) show examples on the Open Images test set across diverse scene types.

Model	Rank	Caption
Global DINO	1	In this image I can see bed with bed sheet and pillows and I can also see table, lamps, something looking like glass and in the background I can see curtains and wall.
	2	This is a picture of a room, in this image in the center there is a bed, on the bed there are pillows and there are lamps, photo frame, tables. On the tables there telephones and some papers, and on the ri...
	3	This picture is clicked inside the room. In this picture, we see the beds and the pillows. We see the blankets in grey color. In between the beds, we see a table on which a telephone and a remote are place...
	4	In this image there are two wooden beds with mattresses and pillows on them. In between the beds on the table there is a telephone and some other objects. Behind the bed there are switches, lamps and curta...
Local DINO	1	In this image I can see a bed. I can see few pillows, towels and blankets on the bed. I can see a lamp on the stool. On the left side I can see few curtains.
	2	In the center of the image we can see beds. On the beds we can see clothes, curtains and some objects. At the bottom we can see shoes on the floor. In the background there is wall.
	3	In this image there is a bed in the room, behind the bed there is a wall, beside the bed there are doors.
	4	This image is taken in the room. In this image there are beds and we can see clothes placed on the beds. There is a television placed on the stand. We can see a table and there is a laptop, lamp and some o...
Global CLIP	1	This is a picture of a room, in this image in the center there is a bed, on the bed there are pillows and there are lamps, photo frame, tables. On the tables there telephones and some papers, and on the ri...
	2	This picture is clicked inside the room. In this picture, we see two beds and the pillows. In between the beds, we see a table on which a remote is placed. On the left side, we see a white color object. In...
	3	In this picture I can see two beds with blankets and pillows on the beds. I can see a lamp on the table and I can see few items on the table. Looks like another lamp on the table on the right side, curtain...
	4	In this picture we can see two beds with pillows and bed sheets on it and these beds and a table on the floor and on this table we can see two lights table lamp, papers, pen and in the background we can se...
Local CLIP	1	In this image I can see two beds with pillows and blankets and I can also see wooden table with drawers, lamp, telephone, some other items and in the background I can see picture frame, floor, door and wal...
	2	In this picture I can see two beds with blankets and pillows on the beds. I can see a lamp on the table and I can see few items on the table. Looks like another lamp on the table on the right side, curtain...
	3	In this picture we can see two beds, there are pillows and bed sheets placed on the beds, in the background we can see a wall, a curtain, a photo frame and a light, there is some text at the right bottom.
	4	This image is taken indoors. In the middle of the image there are two beds with mattress, bed sheets, blankets and pillows on them. On the right side of the image there is a wall. At the left bottom of the...
Original	–	This is an inside view of a room, we can see two beds and there are some pillows on the beds. On the left side, there is a table and we can see the chairs. There is an air conditioner on the wall, we can see the window and curtains.

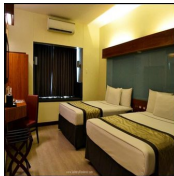


Table A.1: The query image is used to retrieve the captions. None of the retrieved

text is the exact caption of the query image, but still highly relevant captions.

Model	Rank	Caption
Global DINO	1	In this image there are people sitting on wheelchairs and playing basketball, in the background it is blurred.
	2	In this image there are a few people in wheelchairs are playing basketball. Behind them there are a few people sitting in chairs, behind them there is a person performing gymnasium on the ropes. In the bac...
	3	In this image, I can see a group of people sitting on the wheelchairs, which are on the floor. Among them one person is holding a basketball. In the background there is a wall and a railing. In the top lef...
	4	In this image I can see few people playing wheelchair basketball.
Local DINO	1	In the picture I can see the men wearing the sports jersey and they are playing the basketball. I can see the spectators sitting on the chairs.
	2	In this image I can see few people playing basketball. I can see a person sitting on the metal stand, few persons sitting in the stadium and few stairs.
	3	In this image we can see men are playing basketball. In the background, we can see people are sitting and watching the game.
	4	In this image we can see people playing a game. The person in the center is holding a ball. In the background there are people standing.
Global CLIP	1	In this image, I can see a group of people sitting on the wheelchairs, which are on the floor. Among them one person is holding a basketball. In the background there is a wall and a railing. In the top lef...
	2	In this image there are people sitting on wheelchairs and playing basketball, in the background it is blurred.
	3	In this image I can see few people playing wheelchair basketball.
	4	In this image there are a few people in wheelchairs are playing basketball. Behind them there are a few people sitting in chairs, behind them there is a person performing gymnasium on the ropes. In the bac...
Local CLIP	1	In this picture I can see two people wearing medals and sitting on the wheelchairs. I can see a few people standing. I can see a person sitting on the wheelchair on the left side. I can see the roof at the...
	2	In the image there are few people sitting in the wheelchairs. And those wheelchairs are on the road. In the background there are doors which are looking blur.
	3	In this picture we can see a person sitting in wheel chair, holding a bag and talking in phone, back side there is another person pushing the wheel chair.
	4	In this image we can see a person sitting on the racing wheelchair. In the background there are people. At the bottom there is a road.
Original	–	This image is taken indoors. In the middle of the image many people are sitting in the wheelchairs. We can see the floor. On the right side of the image a person is standing on the floor. In the background we can see the boards with text. We can see the chairs. We can see the railings. Many people are sitting on the chairs. We can see the stairs. There is a banner.



Table A.2: The query image

is used to retrieve the captions.

Model	Rank	Caption
Global DINO	1	In this image, we can see a person wearing astronaut suit and he is wearing a hat and the background is dark.
	2	In this image, we can see an astronaut suit in front of the wall.
	3	In front of the image there is a space suit on the display, behind the suit there's a wall.
	4	In the center of the image we can see astronaut suit. In the background we can see wall, curtain and some drawing on the wall. At the top there is light.
	5	In the picture I can see a man wearing a spacesuit.
Local DINO	1	In this image there is a person standing and looking at the right side of the image, behind him there are few people.
	2	In this image there is a person standing.
	3	In this image we can see a person standing on the floor and a stand.
	4	In this image I can see there are few persons walking.
	5	In this image there are people standing, in the background there are clothes.
Global CLIP	1	In this image, we can see a person wearing astronaut suit and he is wearing a hat and the background is dark.
	2	In the picture I can see a man wearing a spacesuit.
	3	In the center of the image we can see astronaut suit. In the background we can see wall, curtain and some drawing on the wall. At the top there is light.
	4	In the image I can see space suits.
	5	In this image there is an astronaut suit which is visible.
Local CLIP	1	In this picture, we see two girls and the boys are standing. They might be exercising. On the right side, we see the legs of two people. At the bottom, we see the floor.
	2	In this image we can see a man and a woman standing.
	3	In this image we can see two persons holding each other and behind them, we can see a woman standing.
	4	In front of the image there is a woman, behind the woman there is a person standing.
	5	In this image we can see there are two women standing with smile, behind them there are few people. The background is dark.
Original	–	In this picture it looks like the cutouts of space suits holding a flag pole with 2 girls standing behind them. In the background, we can see other toys, trees, lights, games etc.,



Table A.3: The query image

is used to retrieve the captions.

Model	Rank	Caption
Global DINO	1	In this image in the center it looks like a toy train and there is a toy railway track, at the bottom it might be a floor.
	2	In this image we can see a scale model, we can see toy trains on the track, which is on the bridge, beneath that there is a tunnel, trees, a building made up of cardboard and a snow. In the background there is a wall and at the top of the image there is a ceiling with lights.
	3	In this image, we can see a train on the track and there are people inside the train. In the background, there are trees, railings, stairs, plants, flowers and there is a rock wall and a shed. On the left, we can see a pole.
	4	In this image, I can see a toy train on a toy railway track and there are few other toy railway tracks. In the top left side of the image, I can see an object on the surface.
Local DINO	1	In this image there is a person standing.
	2	In this image we can see a person standing on the floor and a stand.
	3	In this image there is a person standing and looking at the right side of the image, behind him there are few people.
	4	This image consists of a person. At the bottom, there are clothes to the legs. The person is standing on the floor.
Global CLIP	1	This image consists of miniatures. In this image the we can see the colored background. In the middle of the image we can see the toy houses and buildings. We can see the toy trees and plants. We can see the railings. We can see the toys.
	2	In this image we can see miniature model. There are railway tracks. Also there is a train. And there is a building with windows. Also there is sky.
	3	There is a model of a train in the foreground area of the image and the background is white.
	4	In this image we can see a scale model, we can see toy trains on the track, which is on the bridge, beneath that there is a tunnel, trees, a building made up of cardboard and a snow. In the background there is a wall and at the top of the image there is a ceiling with lights.
Local CLIP	1	In this image there is a person standing and looking at the right side of the image, behind him there are few people.
	2	In the middle of the image a person is standing and watching. Behind him we can see a locomotive.
	3	In this image there is a person standing.
	4	In this image we can see rocks and a person on the land.
Original	–	In this picture we can see a machine on a wooden platform. We can see poles, metal chains, a ladder and few objects. We can see rocks and there are stones on the ground. In the background we can see rock hills and the sky.



Table A.4: The query image

is used to retrieve the captions.

Model	Rank	Caption
Global DINO	1	Here in this picture we can see a group of people standing over a place and we can see all of them are wearing ice skates, gloves and helmet and we can see they are standing on an ice floor and playing ice hockey, as we can see they are holding hockey sticks in their hands.
	2	In the center of the image we can see three people are ice skating and they are in different costumes. And we can see they are holding sticks and they are wearing helmets. In the background there is a wall, transparent glass, banners with some text and barriers. On one of the barriers, we can see an object.
	3	Here in this picture we can see a group of people skating on the ice floor with ice skate under their legs and we can see they are wearing gloves, helmet and holding hockey sticks in their hands and we can see they are playing ice hockey over there and on the right top side we can see a goal post with net present and we can see a person is standing near it with helmet, gloves, knee pads and ice skates and in the bottom we can see the glass walls present.
Local DINO	1	In this picture I can see a group of people wearing ice skating shoes and standing on the ice. There are few people holding hockey sticks. I can see a sports net with poles.
	2	In this image I can see two people with ice-skates and one person holding the stick. These people are on the ice.
	3	In this image we can see the people wearing the helmet and holding the hockey sticks and playing the ice hockey. In the background we can see some person's legs.
	4	In this image, there is a man wearing helmet and seems like he is playing ice hockey. At the top, there are legs of a person.
Global CLIP	1	Here in this picture we can see some people skating on the ice floor present over a place and we can see all of them are wearing ice skates, gloves, helmet and holding a ice hockey stick in their hand and behind them we can see hoarding present and we can also see fencing present and we can see number of people standing and sitting in the stands over there and watching the game.
	2	In this picture we can see a man and a woman wearing ice skates visible on the floor. We can see the lights, stairs and other things. We can see a few people and the dark view in the background.
	3	Here in this picture we can see a group of people standing over a place and we can see all of them are wearing ice skates, gloves and helmet and we can see they are standing on an ice floor and playing ice hockey, as we can see they are holding hockey sticks in their hands.
	4	In this image we can see children doing ice skating on ice. In the back there is a wall.
Local CLIP	1	In this image there is a person standing and looking at the right side of the image, behind him there are few people.
	2	In this image we can see a person standing on the floor and a stand.
	3	In this picture, we see two girls and the boys are standing. They might be exercising. On the right side, we see the legs of two people. At the bottom, we see the floor.
	4	This image consists of a person. At the bottom, there are clothes to the legs. The person is standing on the floor.
Original	—	In the picture I can see the design floor and there are objects on the floor. I can see the lights at the top of the picture. I can see the logos on the wall and there are glass windows on the top left side of the picture.



Table A.5: The query image

is used to retrieve the captions.

MS COCO dataset's Image → Caption Retrieval

Tables A.6, A.7, A.8, A.9, and A.10 show results on the MS COCO validation set.

Model	Rank	Caption
Global DINO	1	A kite being flown in the middle of a beach.
	2	People flying kites on a sandy beach while a bucket sits in the sand.
	3	Kites being used by people on a beach.
	4	A group of people flying kites at the beach
	5	Two people on a beach flying a kite in the air.
Local DINO	1	A kite being flown in the middle of a beach.
	2	People flying kites on a sandy beach while a bucket sits in the sand.
	3	A person standing on top of a beach flying a kite.
	4	Kites being used by people on a beach.
	5	A shot of the blue water with people flying a kite.
Global CLIP	1	A kite being flown in the middle of a beach.
	2	A person standing on top of a beach flying a kite.
	3	A man is flying a kite at the beach.
	4	a man is flying a kite at on the shore at the beach
	5	Kites being used by people on a beach.
Local CLIP	1	A person standing on top of a beach flying a kite.
	2	A kite being flown in the middle of a beach.
	3	People flying kites on a sandy beach while a bucket sits in the sand.
	4	A man is flying a kite at the beach.
	5	A man flying a kite on a beach with people standing around.
Original	–	A man is flying a kite at the beach.

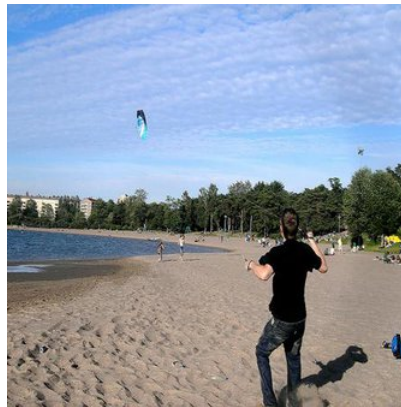


Table A.6: The query image is used to retrieve the captions.

Green color is for the original caption from the dataset.

Model	Rank	Caption
Global DINO	1	A giraffe is walking in some tall grass
	2	A giraffe standing on a grass covered field.
	3	A single giraffe looks over the green brush.
	4	there is a very tall giraffe standing in the wild
	5	a giraffe in a field with trees in the background
Local DINO	1	A giraffe standing on a grass covered field.
	2	A tall giraffe standing on top of a grass covered field.
	3	there is a very tall giraffe standing in the wild
	4	A giraffe is walking in some tall grass
	5	A giraffe standing by a pair of skinny trees.
Global CLIP	1	A giraffe stands near a tree in the wilderness.
	2	A giraffe is walking in some tall grass
	3	A giraffe standing on a grass covered field.
	4	A group of giraffes that are standing in the grass.
	5	there is a very tall giraffe standing in the wild
Local CLIP	1	A tall giraffe standing on top of a grass covered field.
	2	A giraffe standing on a grass covered field.
	3	A giraffe is walking in some tall grass
	4	A single giraffe looks over the green brush.
	5	there is a very tall giraffe standing in the wild
Original	–	A single giraffe looks over the green brush.

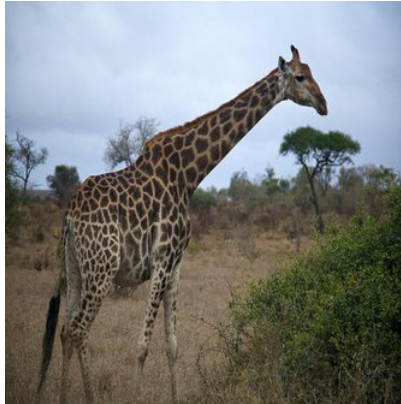


Table A.7: The query image

is used to retrieve the captions.

Green color is for the original caption from the dataset.

Model	Rank	Caption
Global DINO	1	A man riding a surfboard on a wave in the ocean.
	2	A surfer in the ocean trying not to wipeout.
	3	A man on a surfboard riding a wave in the ocean.
	4	a person riding a skate board on a wave
	5	A man is surfing on a small wave.
Local DINO	1	A group of people swimming in the ocean with a surfboard.
	2	there are many surfers that are in the water
	3	Pair of surfers paddling out to open ocean.
	4	A man riding on the back of a surfboard next to kids.
	5	a man on a blue surfboard on top of some rough water
Global CLIP	1	A man riding a surfboard on a wave in the ocean.
	2	A man on a surfboard riding a wave in the ocean.
	3	a person riding a skate board on a wave
	4	a person riding a surf board on a wave
	5	A man on a surfboard, who is riding a wave.
Local CLIP	1	A person on a surfboard in the water.
	2	A para sailor with his board with sail in the surf.
	3	A group of people swimming in the ocean with a surfboard.
	4	A man riding on the back of a surfboard next to kids.
	5	Pair of surfers paddling out to open ocean.
Original	–	a person riding a surf board on a wave

Table A.8: The query image



is used to retrieve the captions.

Green color is for the original caption from the dataset.

Model	Rank	Caption
Global DINO	1	A brown bear walking with rocks in the background.
	2	A large brown bear standing next to a pile of rocks.
	3	A big burly grizzly bear is show with grass in the background.
	4	A majestic bear looks out across a grass plain.
	5	a brown bear is walking away from a river
Local DINO	1	A baby brown bear standing on top of a rock.
	2	A majestic bear looks out across a grass plain.
	3	A brown bear walking with rocks in the background.
	4	A statue of a large brown bear tearing off a cars door.
	5	A brown bear lays down in the woods.
Global CLIP	1	A brown bear walking with rocks in the background.
	2	A large brown bear standing next to a pile of rocks.
	3	A big burly grizzly bear is show with grass in the background.
	4	A baby brown bear standing on top of a rock.
	5	A majestic bear looks out across a grass plain.
Local CLIP	1	A baby brown bear standing on top of a rock.
	2	A majestic bear looks out across a grass plain.
	3	A brown bear walking with rocks in the background.
	4	A big burly grizzly bear is show with grass in the background.
	5	A large brown bear standing next to a pile of rocks.
Original	–	A brown bear walking with rocks in the background.



Table A.9: The query image

is used to retrieve the captions.

Green color is for the original caption from the dataset.

Model	Rank	Caption
Global DINO	1	Person cooking an eggs on a black pot on a stove.
	2	A man pokes his head in front of an oven open to baking cookies.
	3	Belgium waffle loaded with bananas topped with powdered sugar with syrup and more fruit as a garnish.
	4	Several breakfast foods are on top of a refrigerator.
	5	A plate has a waffle, some fruit and ice cream on it.
Local DINO	1	A person is holding a spatula near slices of bread on a stove.
	2	Twp cake pans sitting and cooling on the stove
	3	an image of a man slicing a small pizza
	4	A woman observing something on a kitchen stove.
	5	A man pokes his head in front of an oven open to baking cookies.
Global CLIP	1	A pastry station, with an assortment of fillings and sauces
	2	A young man is working behind a counter.
	3	A group of three chefs preparing food in a kitchen.
	4	A man preparing food in a restaurant kitchen.
	5	The donut robot machine is mechanically making donuts.
Local CLIP	1	A table with many different objects, including a plate of sandwiches.
	2	A pastry station, with an assortment of fillings and sauces
	3	A table topped with plates, bowls and containers of food.
	4	A buffet of casserole dishes on a kitchen counter.
	5	A bunch of items that are on a counter.
Original	–	A pastry station, with an assortment of fillings and sauces



Table A.10: The query image

is used to retrieve the captions.

Green color is for the original caption from the dataset.

A.5.2 Caption → Image Retrieval (In-Distribution)

We evaluate cross-modal alignment through caption-to-image retrieval using SPARC latent representations. Figures [A.16](#), [A.17](#), [A.18](#), and [A.19](#) show retrieval results comparing Global vs Local TopK training configurations.

The four model configurations represent:

- **Global CLIP:** Query caption’s CLIP-text features → Reference database of CLIP-image features (both from Global SPARC)
- **Local CLIP:** Query caption’s CLIP-text features → Reference database of CLIP-image features (both from Local SPARC)
- **Global DINO:** Query caption’s CLIP-text features → Reference database of DINO features (both from Global SPARC)
- **Local DINO:** Query caption’s CLIP-text features → Reference database of DINO features (both from Local SPARC)

Each model shows top-10 retrieved images ranked by cosine similarity in the SPARC latent space.

Open Images dataset’s Image → Caption Retrieval

Figures [A.16](#) and [A.17](#) show results on the Open Images test set using diverse query captions.

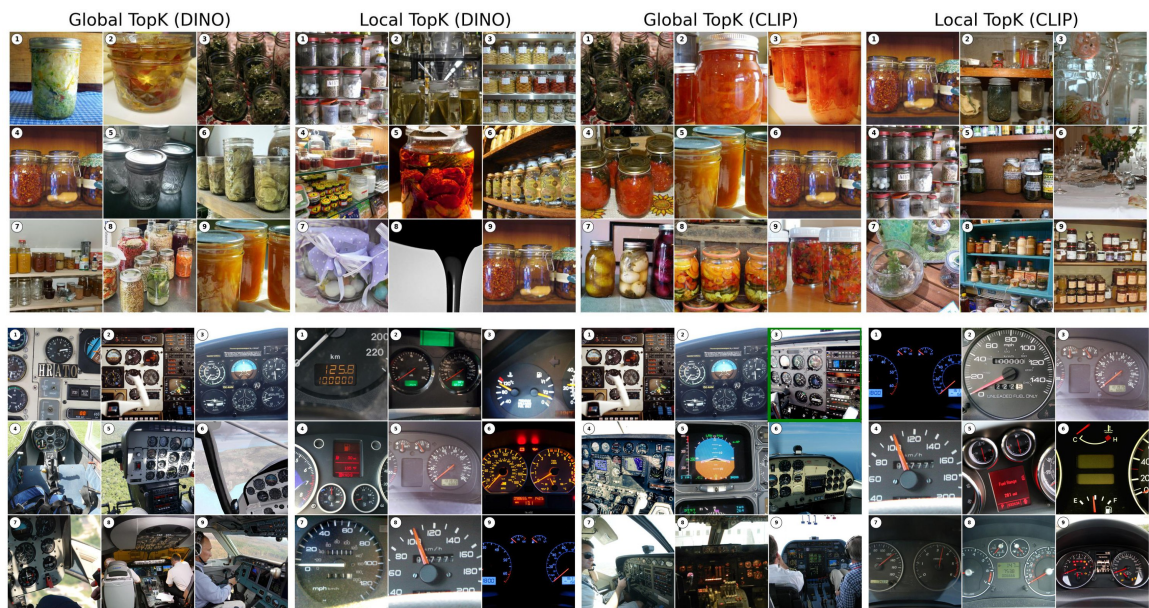


Figure A.16: Images retrieved for captions are (1) "In this picture we can see some food products in the glass jars.", (2) "In this image might be taken in the airplane. In this image we can see the speedometers, knobs and some digital screens." Captions are from Open Images test dataset and images are retrieved from the same split. Green boxes indicate when the corresponding image for a caption is successfully retrieved. The second caption shows such a match (Global TopK with CLIP, 3rd rank).

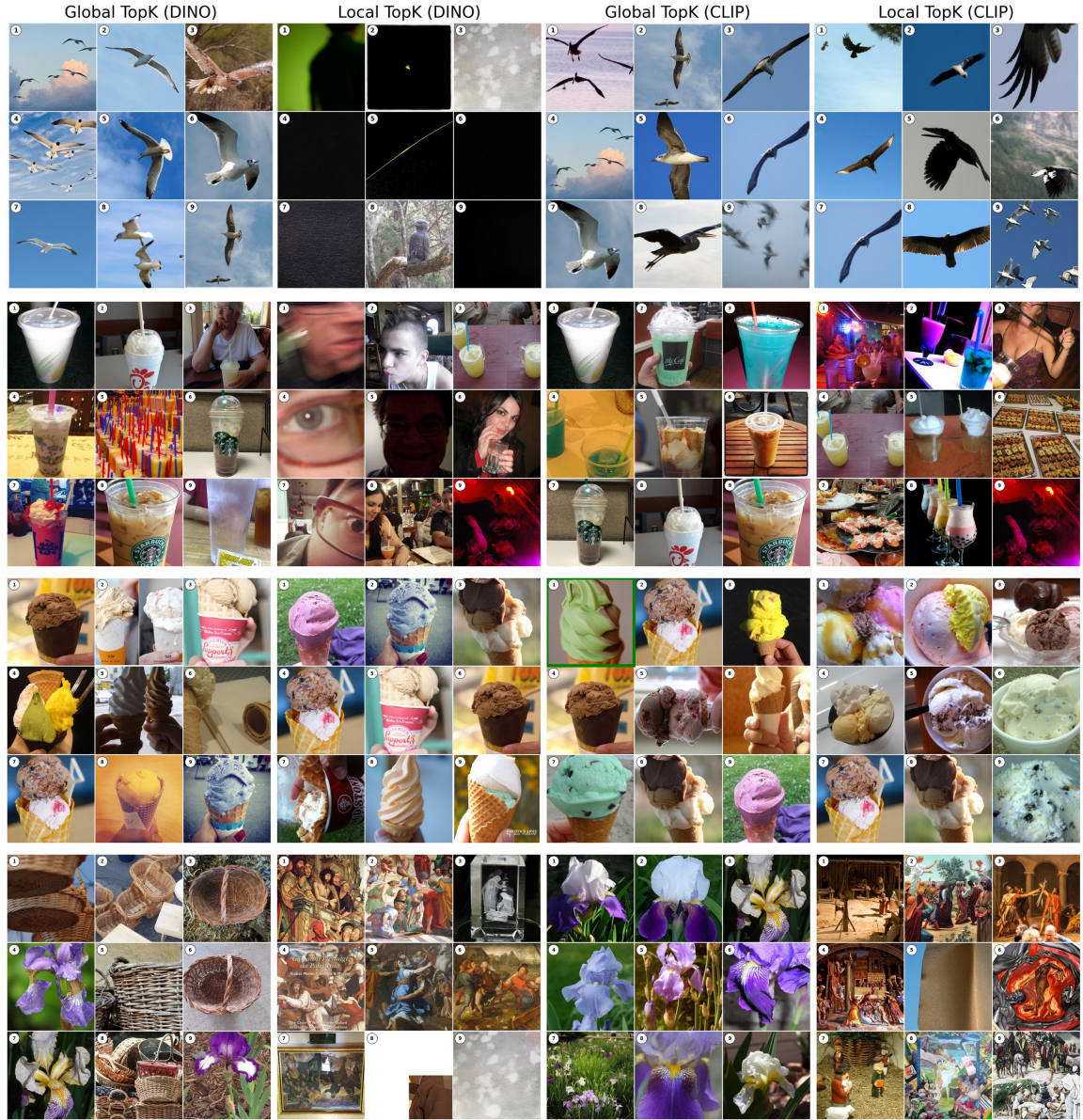


Figure A.17: Images retrieved for captions are (1) "In this image in the center there is one bird flying, and in the background there is sky.", (2) "In this image we can see a table on which some glasses are there in which some food items and straws are there and we can see a pot like structure. On the left side we can see a person hand. In the background we can see some posters and bottles.", (3) "In the picture we can see an ice cream with a green and brown color cream.", and (4) "In this image we can see a wooden basket placed on the ground, we can also see the photo frame on the wall." Captions are from Open Images test dataset and images are retrieved from the same split. Green boxes indicate when the corresponding image for a caption is successfully retrieved. The third caption shows such a match (Global TopK with CLIP, 1st rank).

MS COCO dataset's Image → Caption Retrieval

Figures A.18 and A.19 show results on the MS COCO validation set. We show 1 of the 5 captions available per image in MS COCO in our results.

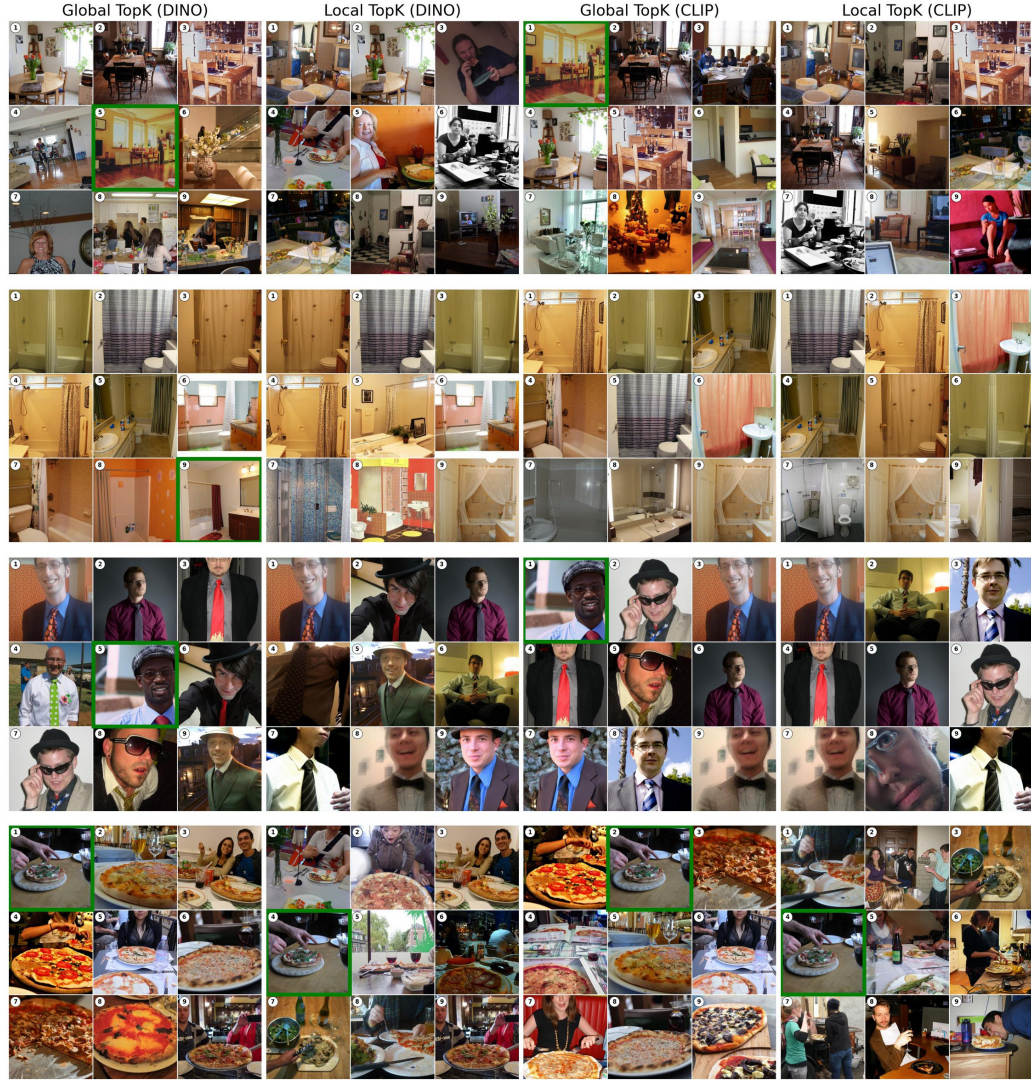


Figure A.18: Images retrieved for captions are (1) "A woman stands in the dining area at the table.", (2) "A shower curtain sits open in an empty and clean bathroom.", (3) "a man in a blue shirt and red tie.", and (4) "These people are going to have pizza and wine." Captions are from Open Images test dataset and images are retrieved from the same split. Green boxes indicate when the corresponding image for a caption is successfully retrieved.

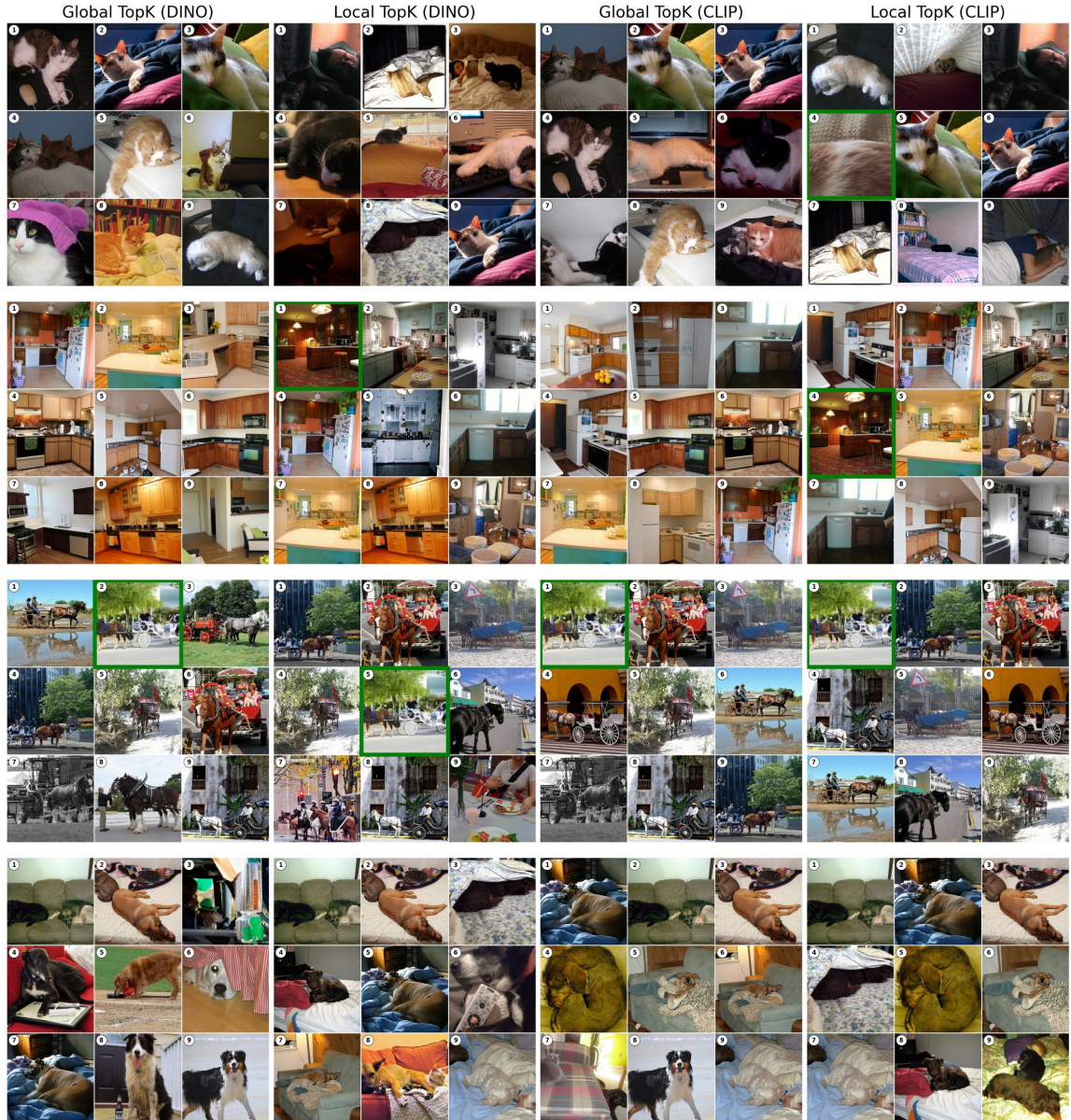


Figure A.19: Images retrieved for captions are (1) "White and orange fur lays on a white blanket.", (2) "A kitchen that has carpeted floors and wooden cabinets.", (3) "A street scene with a horse pulling a white carriage.", and (4) "A large dogs comfortably sleeping on someones bed". Captions are from Open Images test dataset and images are retrieved from the same split. Green boxes indicate when the corresponding image for a caption is successfully retrieved.

A.5.3 Image → Image Retrieval (In-Distribution)

We evaluate cross-model alignment by retrieving images across DINO and CLIP-image encoders using SPARC latent representations. Figures [A.20](#), [A.21](#), [A.22](#), and [A.23](#) show results in a 4-row layout comparing Global vs Local TopK training configurations.

The four rows represent:

- **DINO Global:** Query image’s DINO features → Reference database of CLIP features (both from Global SPARC)
- **DINO Local:** Query image’s DINO features → Reference database of CLIP features (both from Local SPARC)
- **CLIP Global:** Query image’s CLIP features → Reference database of DINO features (both from Global SPARC)
- **CLIP Local:** Query image’s CLIP features → Reference database of DINO features (both from Local SPARC)

Each row shows the query image (left) followed by top-10 retrieved images.

Open Images dataset’s Image → Image Retrieval

Figures [A.20](#), [A.21](#), and [A.22](#) show cross-model retrieval results on the Open Images test set.

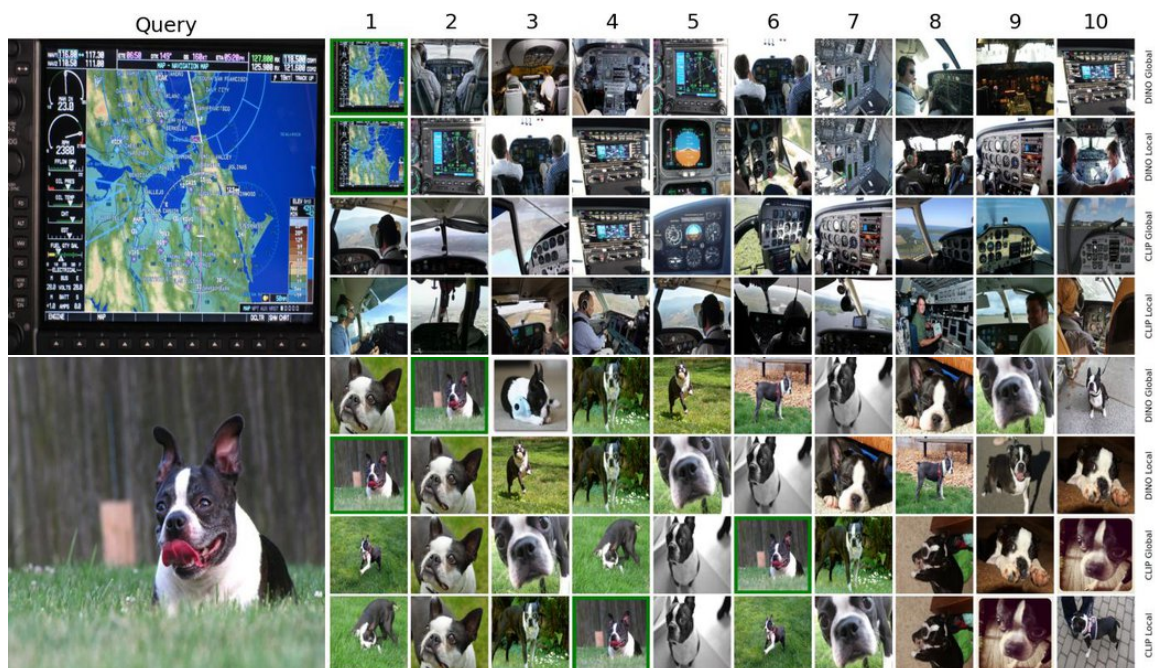


Figure A.20: Cross-model image retrieval results. Each image shows a 4-row layout comparing query stream and reference database combinations. Query image (left) with top-10 retrieved images from reference database (right). All models trained with $\lambda = 1$ on Open Images training set. The retrieval is done on the test set of Open Images. Green border is used to show the exact match of the query image was found.

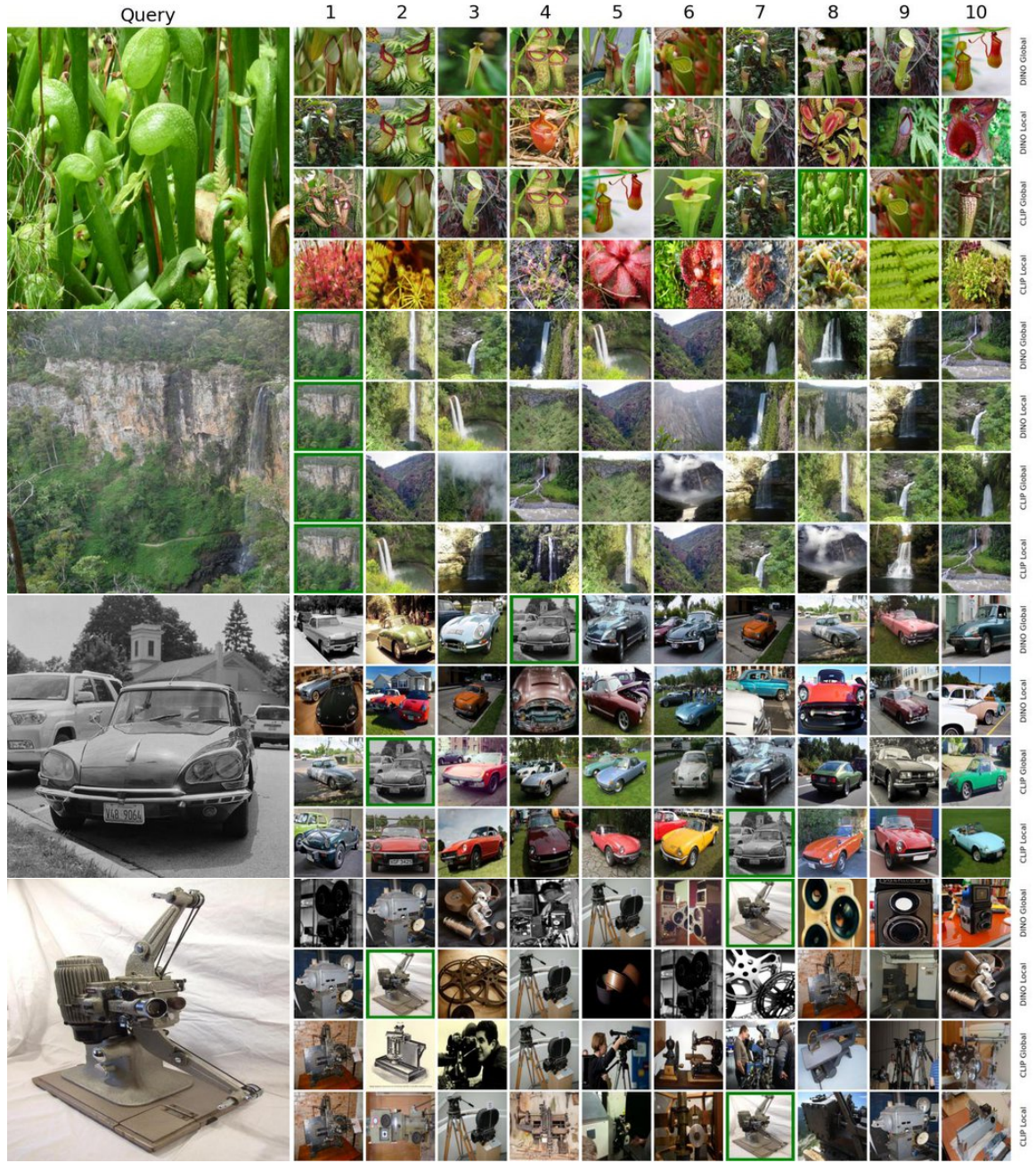


Figure A.21: Cross-model image retrieval results. Each image shows a 4-row layout comparing query stream and reference database combinations. Query image (left) with top-10 retrieved images from reference database (right). All models trained with $\lambda = 1$ on Open Images training set. The retrieval is done on the test set of Open Images. Green border is used to show the exact match of the query image was found.

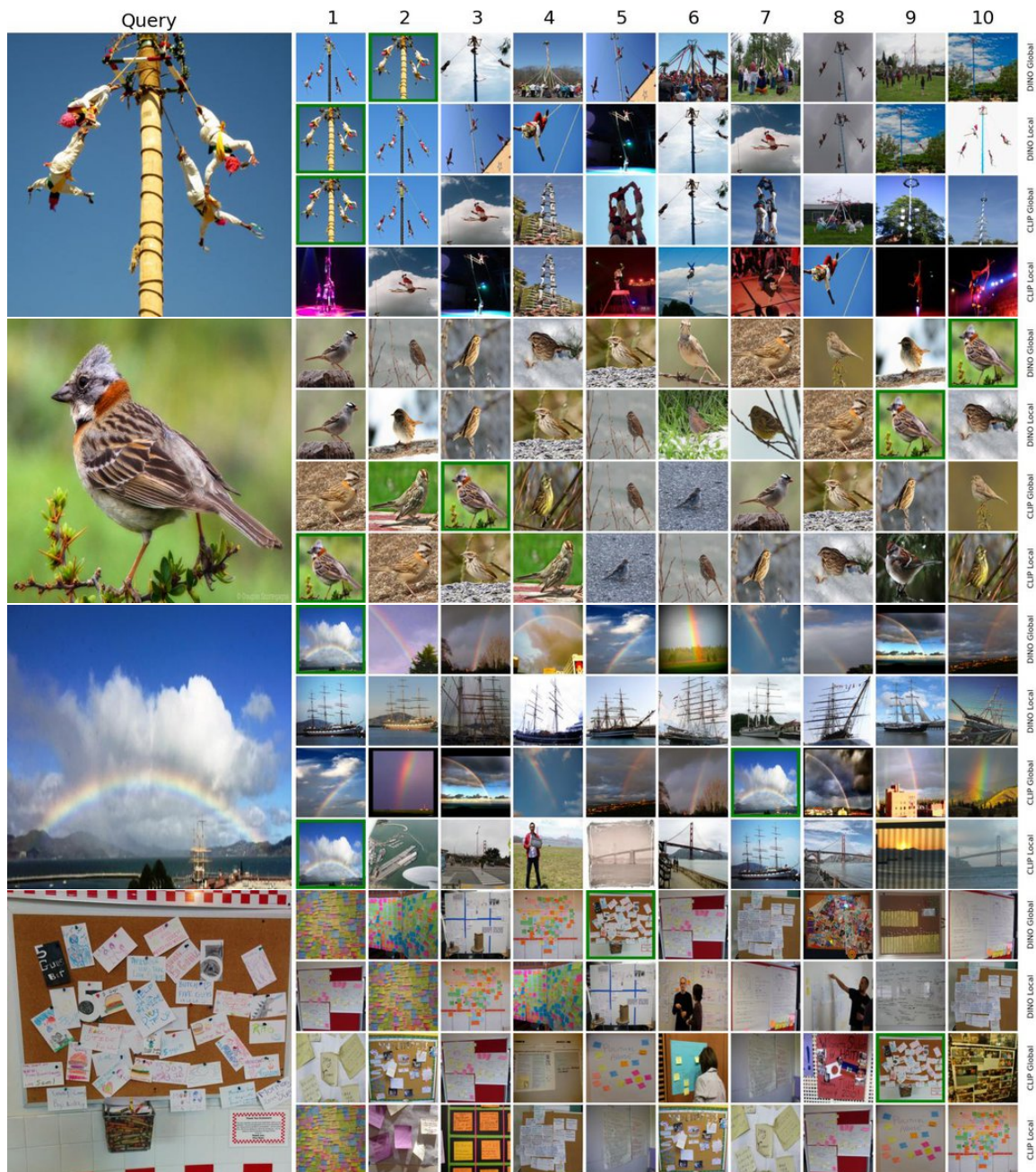


Figure A.22: Cross-model image retrieval results. Each image shows a 4-row layout comparing query stream and reference database combinations. Query image (left) with top-10 retrieved images from reference database (right). All models trained with $\lambda = 1$ on Open Images training set. The retrieval is done on the test set of Open Images. Green border is used to show the exact match of the query image was found.

MS COCO dataset’s Image → Image Retrieval Figure [A.23](#) shows cross-model image-to-image retrieval results on the MS COCO validation set.

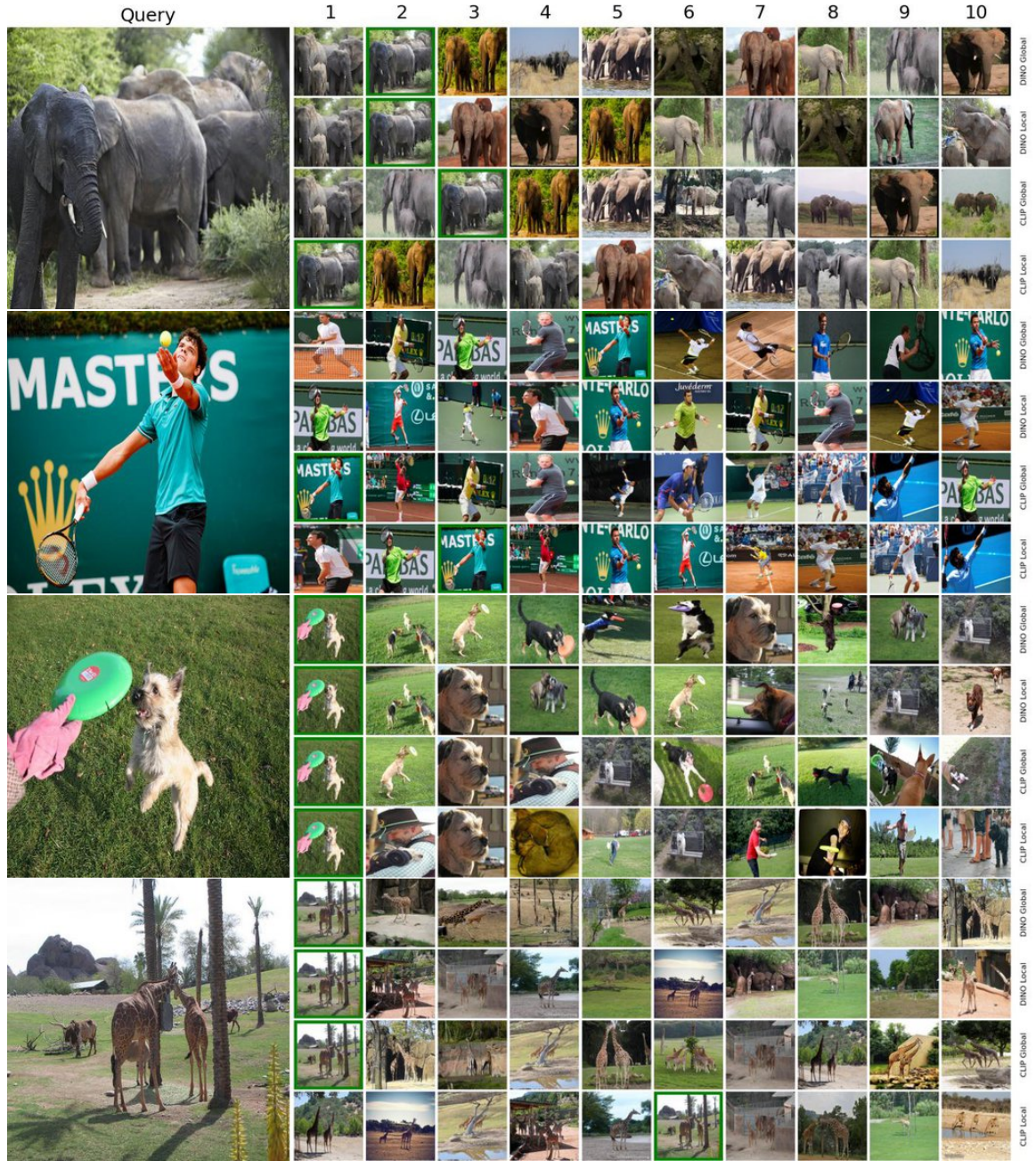


Figure A.23: Cross-model image retrieval results. Each image shows a 4-row layout comparing query stream and reference database combinations. Query image (left) with top-10 retrieved images from reference database (right). All models trained with $\lambda = 1$ on MS COCO training set. The retrieval is done on the validation set of MS COCO. Green border is used to show the exact match of the query image was found.

A.5.4 External Image \rightarrow Caption (OOD)

We evaluate SPARC’s out-of-distribution generalization using external images not present in the Open Images dataset, following the same methodology as Section A.5.1. Tables A.11, A.12, A.13, A.14, A.15, A.16, and A.17 show top-5 retrieved captions for each model configuration using external query images with the Open Images test set as the reference database.

Model	Rank	Retrieved Caption
Global DINO	1	In this image we can see a laptop with a screen and keys. At the bottom of the image there is a surface. On the surface we can see reflections. On the image there is a watermark.
	2	In this image we can see black and white picture of a laptop, we can also see some pictures on the screen.
	3	In this image there is a poster. There is a screen, sound speakers, laptop on the devices. Bottom of the image there is some text. Background is in grey color. Top of the image there is some text.
	4	In this image we can see a laptop containing some text on its screen.
	5	In this picture we can see keypad of laptop.
Local DINO	1	In this image we can see a laptop.
	2	In this picture I can observe black color joystick on the laptop. There is an apple logo on the laptop. The background is in white color.
	3	In this image we can see a laptop containing some text on its screen.
	4	There is a white color laptop present on the cloth as we can see in the middle of this image. There is one device connected to the laptop. It is dark in the background.
	5	In the center of the image we can see the text. In the background of the image we can see a laptop. On the laptop we can see the objects.
Global CLIP	1	This image is taken indoors. In the background there is a wall. In the middle of the image there are two laptops on the table.
	2	In this picture we can see laptop, keys and screen. In the background of the image it is blurry.
	3	In this image there is a laptop on the table. There are letters, numbers, symbols on the keyboard. There is some text at the top of the image.
	4	In this picture we can see a laptop with keys and on this laptop screen we can see text, symbols and buttons.
	5	In this picture we can see a laptop. We can see a person, numbers, a symbol and a few things on the screen of this laptop.
Local CLIP	1	In this picture it looks like a laptop. We see the laptop keyboard which has the alphabet and the number keys. We see the text written on the laptop and we see the buttons. At the bottom, we see the touchpad on the laptop.
	2	In this picture we can see a laptop with keys and on this laptop screen we can see text, symbols and buttons.
	3	In the image there is a laptop on a cloth and in front of the laptop there is a teddy bear and it seems like both the things are kept on a sofa and in the background there is a wall.
	4	In this picture we can see a laptop. We can see a person, numbers, a symbol and a few things on the screen of this laptop.
	5	In this image, I can see a laptop. On which I can see keys having letters, numbers, symbols and some text.



Table A.11: The query image

is used to retrieve the captions.

Model	Rank	Retrieved Caption
Global DINO	1	In this image we can see a laptop with a screen and keys. At the bottom of the image there is a surface. On the surface we can see reflections. On the image there is a watermark.
	2	In this image we can see black and white picture of a laptop, we can also see some pictures on the screen.
	3	In this image there is a poster. There is a screen, sound speakers, laptop on the devices. Bottom of the image there is some text. Background is in grey color. Top of the image there is some text.
	4	In this image we can see a laptop containing some text on its screen.
	5	In this picture we can see keypad of laptop.
Local DINO	1	In this image we can see a laptop.
	2	In this picture I can observe black color joystick on the laptop. There is an apple logo on the laptop. The background is in white color.
	3	In this image we can see a laptop containing some text on its screen.
	4	There is a white color laptop present on the cloth as we can see in the middle of this image. There is one device connected to the laptop. It is dark in the background.
	5	In the center of the image we can see the text. In the background of the image we can see a laptop. On the laptop we can see the objects.
Global CLIP	1	This image is taken indoors. In the background there is a wall. In the middle of the image there are two laptops on the table.
	2	In this picture we can see laptop, keys and screen. In the background of the image it is blurry.
	3	In this image there is a laptop on the table. There are letters, numbers, symbols on the keyboard. There is some text at the top of the image.
	4	In this picture we can see a laptop with keys and on this laptop screen we can see text, symbols and buttons.
	5	In this picture we can see a laptop. We can see a person, numbers, a symbol and a few things on the screen of this laptop.
Local CLIP	1	In this picture it looks like a laptop. We see the laptop keyboard which has the alphabet and the number keys. We see the text written on the laptop and we see the buttons. At the bottom, we see the touchpad on the laptop.
	2	In this picture we can see a laptop with keys and on this laptop screen we can see text, symbols and buttons.
	3	In the image there is a laptop on a cloth and in front of the laptop there is a teddy bear and it seems like both the things are kept on a sofa and in the background there is a wall.
	4	In this picture we can see a laptop. We can see a person, numbers, a symbol and a few things on the screen of this laptop.
	5	In this image, I can see a laptop. On which I can see keys having letters, numbers, symbols and some text.

Table A.12: The query image



is used to retrieve the captions.

Model	Rank	Retrieved Caption
Global DINO	1	In this picture we can see a few people, snowy mountains and hills. We can see other things and the cloudy sky in the background.
	2	In the picture we can see the mountains covered with the snow and behind it, we can see the sky with clouds.
	3	In this picture we can see the man wearing a black jacket and standing. In the front we can see some stones. Behind we can see mountains and some snow. On the top we can see the sky and clouds.
	4	In this picture I can see few people. There are rocks. I can see snowy mountains, and in the background there is the sky.
	5	In this image we can see mountains with snow. In the background there is sky.
Local DINO	1	In this picture I can see mountains and I can see snow and a blue cloudy sky. I can see text at the bottom left corner of the picture and looks like a cross symbol on the mountain.
	2	In the image we can see the person standing, wearing clothes, gloves, shoes and spectacles, here we can see the stones, snow, mountains and the sky.
	3	In this picture, we see the man in the jacket is standing. He is wearing a red cap, spectacles and the gloves. On the left side, we see the snow. In the background, we see the hills. These hills are covered with the snow.
	4	In this image we can see a man in blue color jacket is sitting on a snow covered ground and he is also wearing a white color cap and a gaggle as well. We can also see a bag and a stick on the snow covered ground. In background we can see a mountain and a clear blue sky.
	5	In this picture I can see there is a person and he is wearing a coat and a bag pack, there are a few mountains in the background and they are covered with rocks and snow.
Global CLIP	1	In this picture I can see mountains and I can see snow and a blue cloudy sky. I can see text at the bottom left corner of the picture and looks like a cross symbol on the mountain.
	2	In the foreground we can see rocks, mountain, soil and people. In the middle of the image there are mountains. In the background it is the sky.
	3	In this image I can see a person standing on the mountain, he is wearing a bag. There are few rocks on the mountain, I can see there are few mountains in the background and the sky is clear.
	4	In this image we can see so many rocks and snow on the mountain. In background we can see some more mountains and a clear blue sky.
	5	In this picture I can see few people. There are rocks. I can see snowy mountains, and in the background there is the sky.
Local CLIP	1	In this image I can see some people on the mountain, and there is snow on the mountain and the sky is blue.
	2	In this picture we can see a few people, snowy mountains and hills. We can see other things and the cloudy sky in the background.
	3	In this picture I can see a person standing in a foreground of the image and there are a few rocks visible, in the background there is a mountain visible and there are few trees and there is snow covered on the mountain and I can see the sky.
	4	In this image in the center there are a group of people who are standing, and they are wearing bags. And at the bottom there is snow and one stick is there, and in the background there are mountains and at the top of the image there is sky.
	5	In the image there are few people sitting on the rocks. And there is a person standing on the rocks. In the background there is a hill covered with snow. And also there are few hills in the background. At the top of the image there is sky.

Table A.13: The query image



is used to retrieve the captions.

Model	Rank	Retrieved Caption
Global DINO	1	In this picture I can see a sea otter in the water and looks like a rock at the top left corner.
	2	In the middle of this image I can see a sea otter in the water. At the bottom, I can see the sticks and leaves in the water.
	3	In this picture, we see a sea otter is swimming in the water. In the background, we see the water and this water might be in the swimming pool.
	4	At the bottom of the picture, we see the rocks. In front of the picture, we see a walrus is sleeping on the rock. Behind that, we see water and this water might be in the pond. There are rocks and a walrus in the background.
	5	In this image we can see a sea lion in the water.
Local DINO	1	In this image we can see a paper. On the paper we can see painting of birds on branch. Also we can see leaves. And there is text on the paper.
	2	In this image, we can see painting of parrots on the branch and some text at the bottom.
	3	In this image, we can see a water animal picture on a cream surface. At the top and bottom of the image, we can see text.
	4	In this image we can see a red color crab with two sticks on a paper on which something is written.
	5	In this picture I can see the bactrian camel in the foreground. It is looking like the green grass, plants in the background. It is looking like the fence on the right side.
Global CLIP	1	In this image we can see a sea lion in the water.
	2	In the image we can see water, on the water we can see some rafts. On the rafts we can see some seals.
	3	In this image I can see two sea lions in the water.
	4	In this image we can see sea lion in the water.
	5	In this image we can see sea lion in the water.
Local CLIP	1	In this image I can see the hippopotamus and the rock in the water.
	2	In this image we can see hippopotamus. Also we can see water. In the back we can see stones.
	3	In this picture, we see a hippopotamus is in the water. It might be a pool. At the bottom, we see a wall. In the right top, we see the railing and a wall. We see the rods or the stands in the pool.
	4	In this image we can see hippopotamus in water and ground. In the background we can see trees.
	5	In this image I can see an animal in the water, looks like hippopotamus.



Table A.14: The query image is used to retrieve the captions.

Model	Rank	Retrieved Caption
Global DINO	1	In this image there is ground at the bottom. There are rocks in the foreground. And there is greenery in the background. And there is sky at the top.
	2	In this image there are few rocks, grass, a river, tree and in the background there is the sky.
	3	In this picture we can see rocks, trees and in the background we can see the sky with clouds.
	4	In this image we can see rocks. On the ground there is grass. In the background there is sky.
	5	In this image we can see rocks. Also there are people. In the background there is sky.
Local DINO	1	This image consists of stones, rocks, grass, a group of people, trees and the sky.
	2	In this image we can see stones. Also we can see grass on the ground. There are trees. In the background there is sky with clouds.
	3	In this picture we can see stones and here we can see the grass and trees on the ground. In the background we can see the sky.
	4	In this image we can see rocks. On the ground there is grass. In the background there is sky.
	5	In this picture, we see the stones, rocks and the grass. In the background, we see the stones and the rocks.
Global CLIP	1	In this image we can see stones. Also we can see grass on the ground. There are trees. In the background there is sky with clouds.
	2	In this image we can see rocks. Also there is water. On the left side we can see grass on the ground. In the background there is sky with clouds.
	3	In this picture we can see stones and here we can see the grass and trees on the ground. In the background we can see the sky.
	4	In this image we can see in front there are many rocks, trees, at the back there are hills, mountains, the sky is at the top.
	5	There are stones and grassland in the foreground area of the image, there are people, trees, grassland and the sky in the background.
Local CLIP	1	In this image we can see rocks. On the ground there is grass. In the background there is sky.
	2	In this picture, we see the stones, rocks and the grass. In the background, we see the stones and the rocks.
	3	This image consists of stones, rocks, grass, a group of people, trees and the sky.
	4	There are stones and grassland in the foreground area of the image, there are people, trees, grassland and the sky in the background.
	5	In this picture we can see stones and here we can see the grass and trees on the ground. In the background we can see the sky.



Table A.15: The query image

is used to retrieve the captions.

Model	Rank	Retrieved Caption
Global DINO	1	In this image in the foreground there is a tree, and at the bottom there is a river and in the background there are hills and trees. And at the top there is sky.
	2	In this image, there are mountains and clouds.
	3	In this picture we can see hills in the background, it looks like snow at the bottom, we can see the sky at the top of the picture, there are clouds in the middle.
	4	In this image, there is a sewing machine. Under this sewing machine, there is a sheet. On the right side of this image, there is a gray color sheet. And the background of this image is dark in color.
	5	In this image, there are few electronic devices on a desk and also there are drawers at the bottom. Beside, there is a woman sitting on the chair on the floor and operating a computer. In the background, there are glass walls, few lights to the ceiling, a chair, dustbin, few electronic devices on the desk and also there are few drawers on the right.
Local DINO	1	In this image there is a dog on the rock. Behind the dog there are plants and trees. In the background of the image there are mountains. At the top of the image there are clouds in the sky.
	2	In this image I can see a person standing on the mountain, he is wearing a bag. There are few rocks on the mountain, I can see there are few mountains in the background and the sky is clear.
	3	In this image we can see a person sitting on a rock, in front of the person there is a water flow and there is a dog. In the background there is a snow on the mountains and the sky.
	4	In the image there are cats sitting on the stones. At the top of the image there is water.
	5	In this picture there are dogs and we can see rocks, trees and water. In the background of the image we can see hills and sky with clouds.
Global CLIP	1	In this image there is a dog on a couch which is on the floor. Background is blurry.
	2	In the image I can see the picture of a dog which is on the seat.
	3	In the image there is a dog on a couch. And also there is a pillow and a towel.
	4	In this image there are two dogs on a sofa, on that dogs there is a blanket, in the background there is a wall.
	5	In this picture, we see a dog is on a sofa or on a bed. The dog is in black and white color. The leash of the dog is in red color. In the background, we see a bed sheet or a cloth in brown color.
Local CLIP	1	In the image there is a dog walking on the rock surface, behind the dog there are huge rocks.
	2	In this image we can see a person sitting on a rock, in front of the person there is a water flow and there is a dog. In the background there is a snow on the mountains and the sky.
	3	In this image, I see a group of dogs standing on a stone field with belts around their necks and I see couple of persons standing beside them.
	4	At the bottom of the image there is a dog in the water. In the background there are rocks. And also there are few stones in the water.
	5	In this picture there are dogs and we can see rocks, trees and water. In the background of the image we can see hills and sky with clouds.

Table A.16: The query image



is used to retrieve the captions.

Model	Rank	Retrieved Caption
Global DINO	1	In this image I can see two umbrellas with lot of colors like pink, blue, yellow, orange, red, green, and the sky is cloudy.
	2	In this image I can see different color of umbrellas. In the background I can see clouds in the sky.
	3	This image consist an umbrella. In the middle, we can see a pole. The background is blue in color.
	4	This image consists of an umbrella along with a metal rod. At the top, there is sky.
	5	In this picture we can see a partial part of a colorful umbrella. We can see the top view of an umbrella.
Local DINO	1	In this image we can see a blue color umbrella and on the umbrella we can see some different paintings in different colors and in the background we can see a white color iron and we can see the tip of the umbrella is in red color.
	2	In the middle of the image we can see an umbrella on the dried grass.
	3	In this image, there is a black and red umbrella on the white surface. Behind it, there is a white wall.
	4	In this picture there is a woman holding the umbrella. At the back there is a wall. At the bottom there is a floor.
	5	In this image we can see a man holding an umbrella.
Global CLIP	1	In this image we can see a blue color umbrella and on the umbrella we can see some different paintings in different colors and in the background we can see a white color iron and we can see the tip of the umbrella is in red color.
	2	This image consist an umbrella. In the middle, we can see a pole. The background is blue in color.
	3	In this image, I can see an umbrella with colorful design and there is a thread. In the background, I can see glass windows. In the bottom left corner of the image, there is an object.
	4	In this image I can see two umbrellas with lot of colors like pink, blue, yellow, orange, red, green, and the sky is cloudy.
	5	In this image, we can see an umbrella.
Local CLIP	1	In this image we can see a man holding an umbrella.
	2	In this picture there is a woman holding the umbrella. At the back there is a wall. At the bottom there is a floor.
	3	In this image, we can see an umbrella.
	4	In the middle of the image we can see an umbrella on the dried grass.
	5	In this image, in the foreground we can see a person wearing a costume and hold an umbrella. On the right side, we can see a person. Background of the image is blurred.



Table A.17: The query image is used to retrieve the captions.

A.5.5 Free-Form Caption → Image (OOD)

We evaluate SPARC’s capability with free-form captions not present in the Open Images dataset, following the same methodology as Section A.5.2. Figures A.24, A.25, and A.26 show retrieval results using simple descriptive captions to query the Open Images test database.

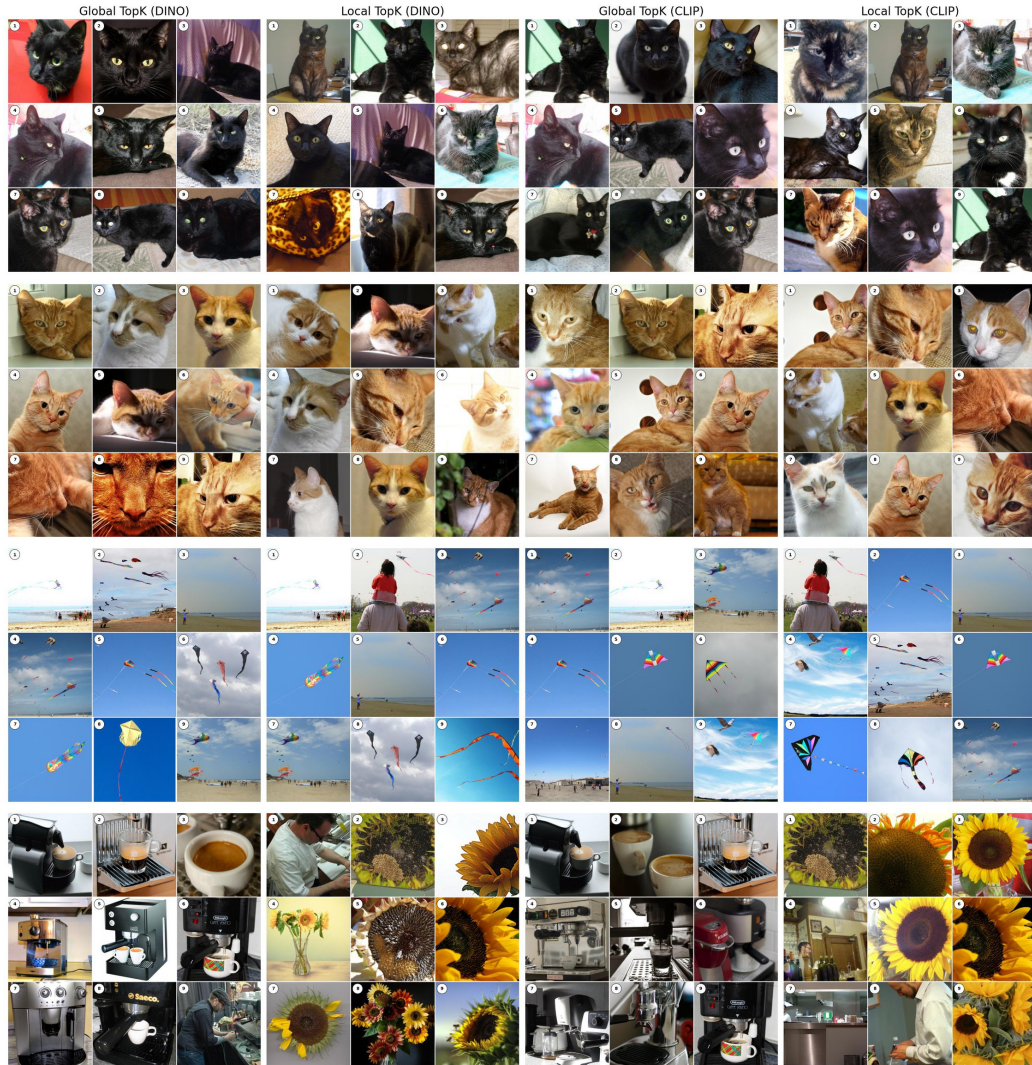


Figure A.24: Images retrieved for captions: (1) "Black cat", (2) "Orange cat", (3) "A child flying a red kite on a sunny beach", (4) "A barista making coffee behind a counter." The images are retrieved from the test set of the Open Images dataset, but the captions are not part of the dataset.



Figure A.25: Images retrieved for captions: (1) "People crossing a busy city street in the rain", (2) "A passenger airplane flying in a clear blue sky", (3) "A dog running through a grassy field", (4) "A man reading a newspaper at a bus stop." The images are retrieved from the test set of the Open Images dataset, but the captions are not part of the dataset.

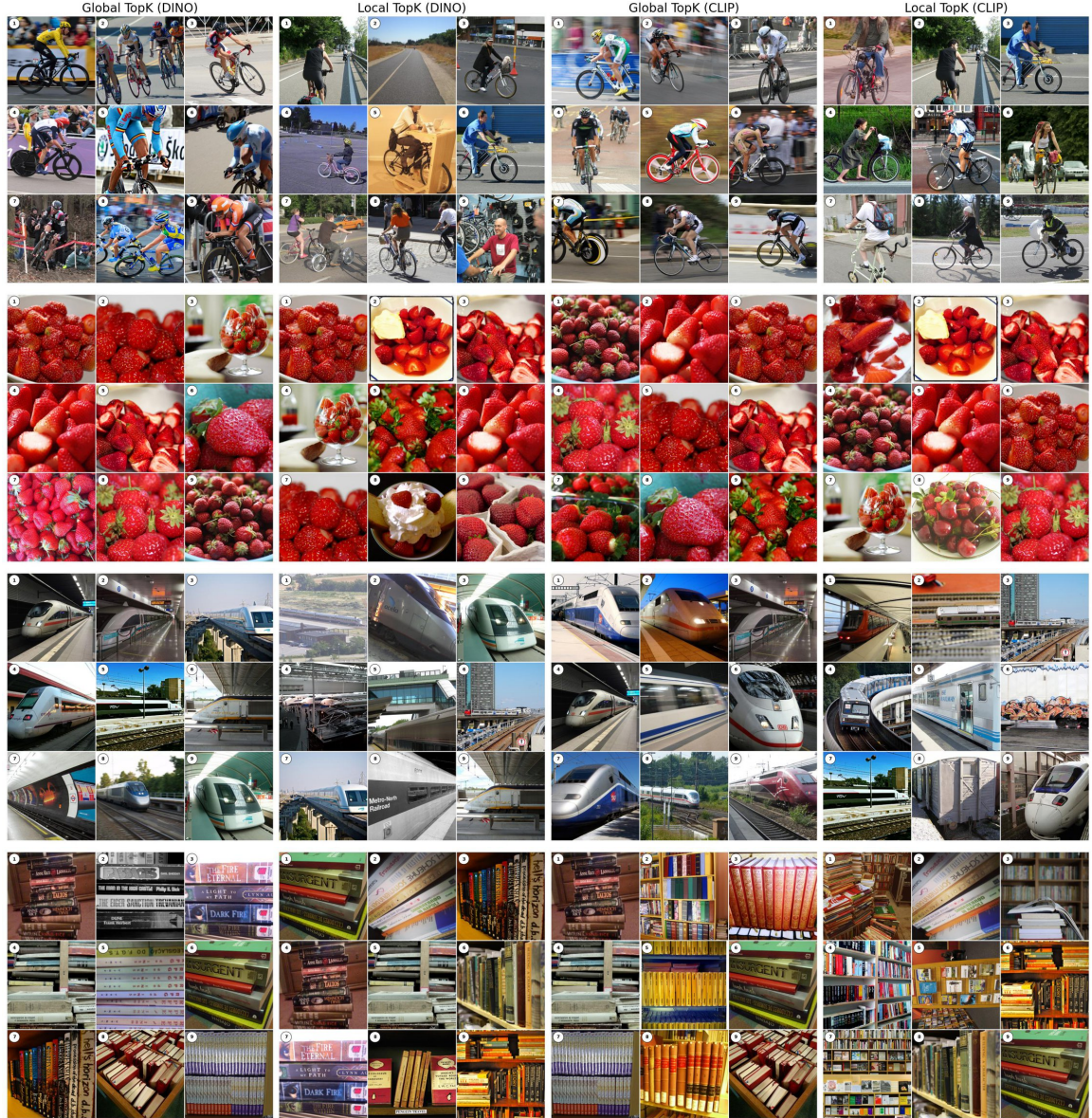


Figure A.26: Images retrieved for captions: (1) "A person riding a bicycle on a country road", (2) "A bowl of fresh strawberries on a kitchen counter", (3) "A train arriving at an underground metro station", (4) "Books stacked on a wooden desk." The images are retrieved from the test set of the Open Images dataset, but the captions are not part of the dataset.

A.5.6 External Image→Image (OOD)

We evaluate SPARC’s out-of-distribution generalization using external images not present in the Open Images dataset for cross-model image retrieval. Figures [A.27](#), [A.28](#), and [A.29](#) show cross-model retrieval results using external query images with the Open Images test set as the reference database.

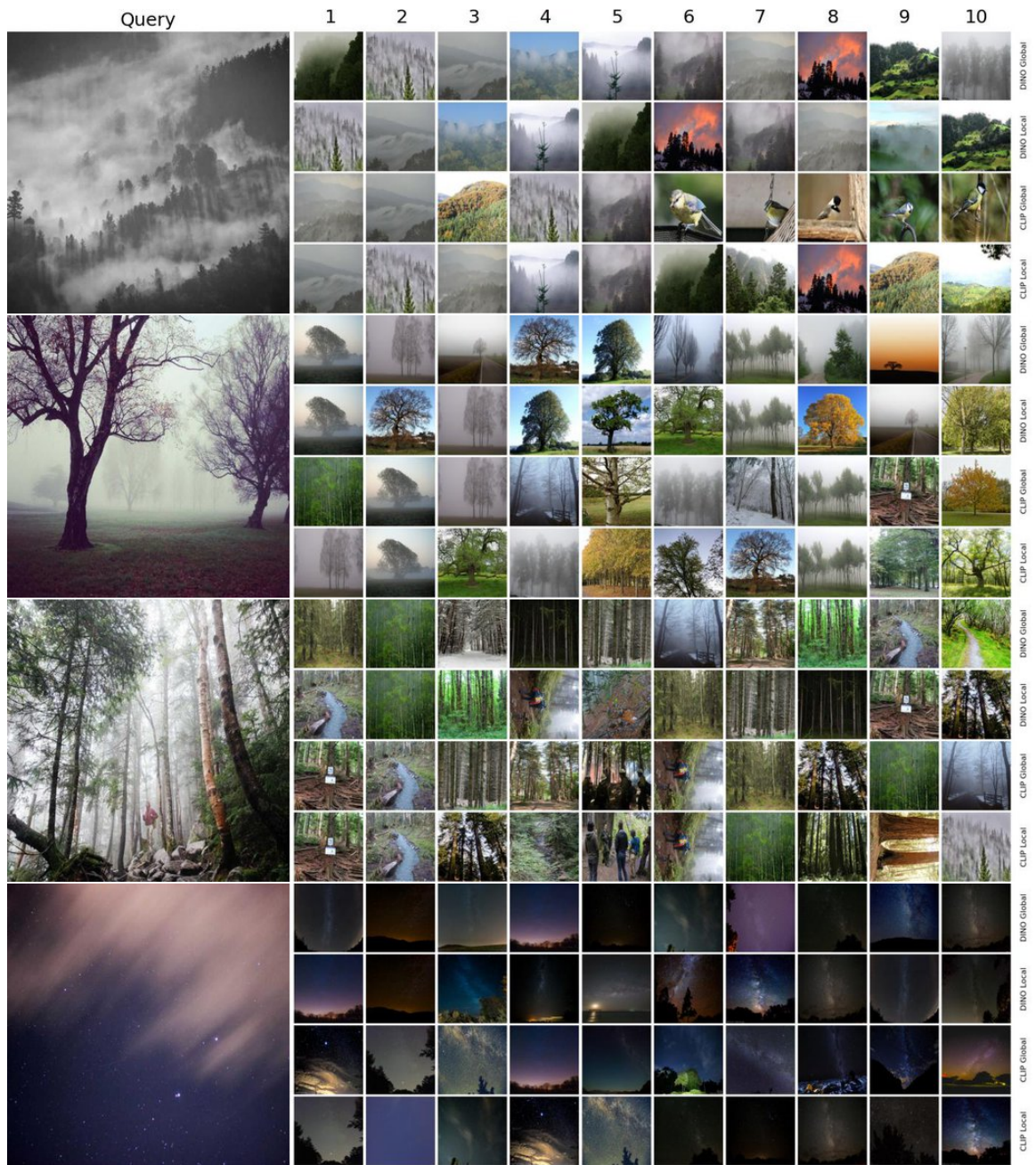


Figure A.27: Cross-model image retrieval results using OOD query images. The references are from the Open Images test set.

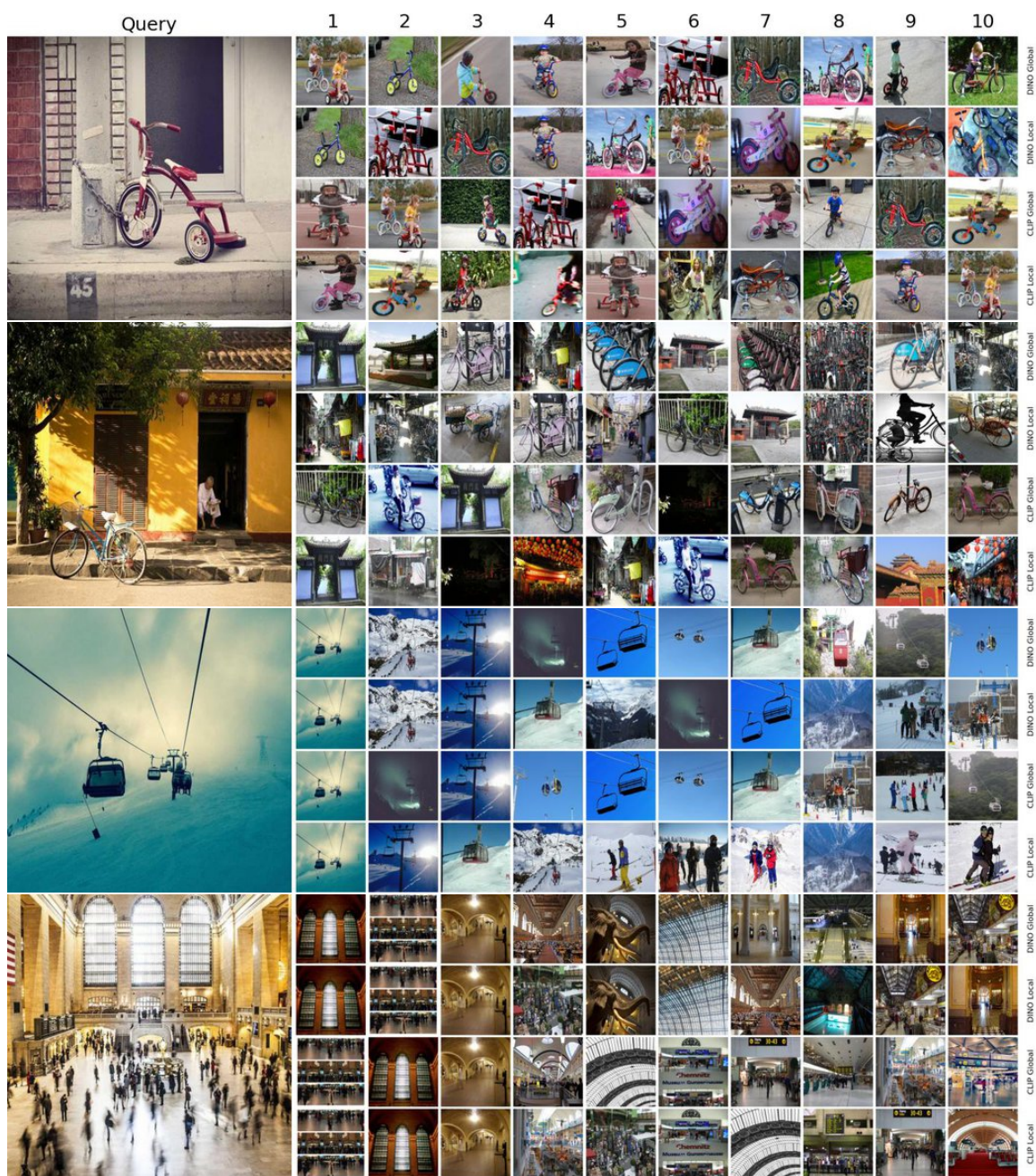


Figure A.28: Cross-model image retrieval results using OOD query images. The references are from the Open Images test set.

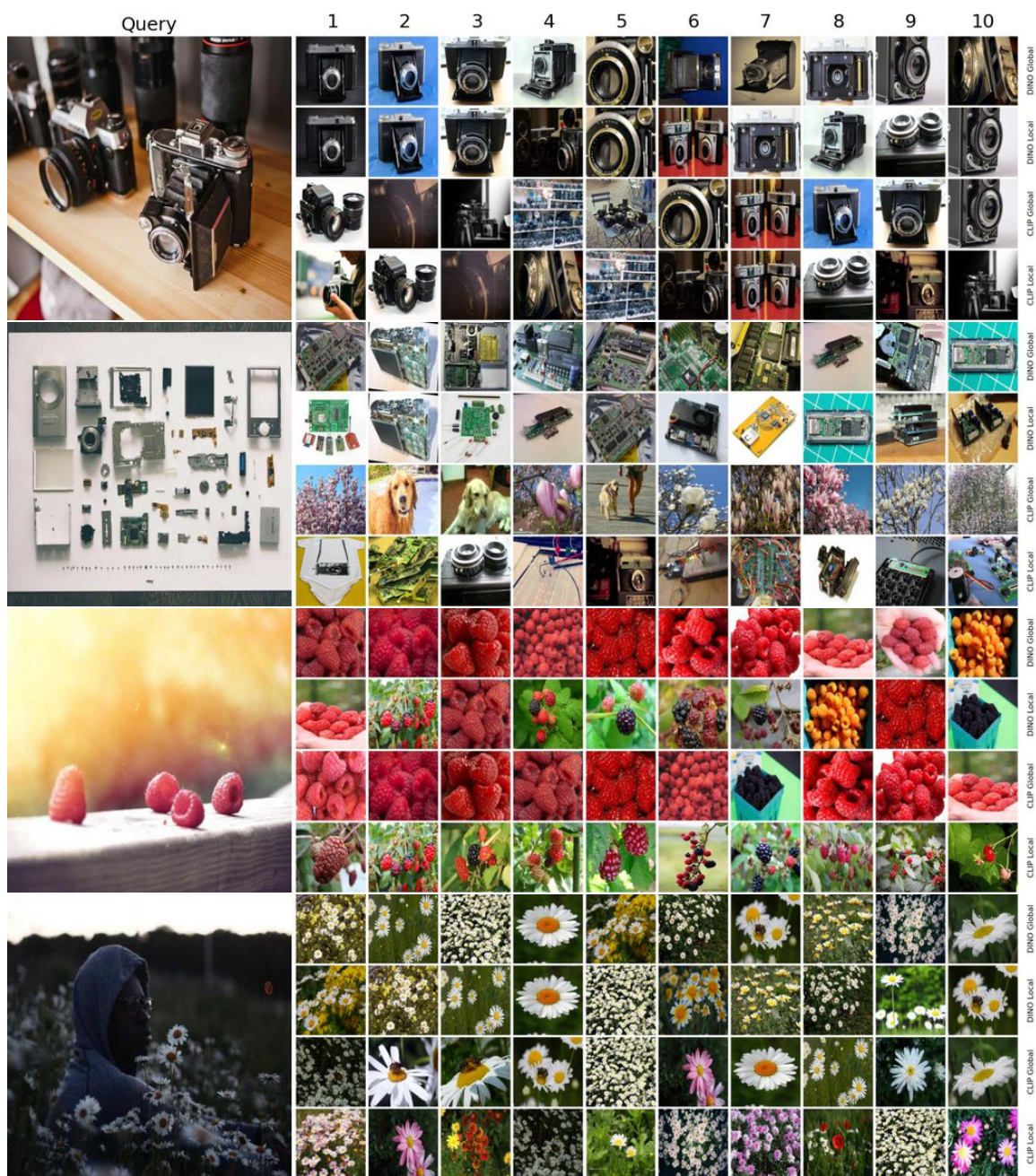


Figure A.29: Cross-model image retrieval results using OOD query images. The references are from the Open Images test set.

References

- Afrin, H., Larson, N. B., Fatemi, M., & Alizad, A. (2023). Deep learning in different ultrasound methods for breast cancer, from diagnosis to prognosis: current trends, challenges, and an analysis. *Cancers*, 15(12), 3139.
- Agrawal, P., Carreira, J., & Malik, J. (2015). Learning to see by moving. In *2015 IEEE International Conference on Computer Vision (ICCV)* (p. 37-45). doi: 10.1109/ICCV.2015.13
- Aharon, M., Elad, M., & Bruckstein, A. (2006). K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11), 4311-4322. doi: 10.1109/TSP.2006.881199
- Al-Dhabyani, W., Gomaa, M., Khaled, H., & Fahmy, A. (2020). Dataset of breast ultrasound images. *Data in brief*, 28, 104863.
- Ali, M. D., Saleem, A., Elahi, H., Khan, M. A., Khan, M. I., Yaqoob, M. M., ... Al-Rasheed, A. (2023). Breast cancer classification through meta-learning ensemble technique using convolution neural networks. *Diagnostics*, 13(13), 2242.
- Asano, Y. M., Rupprecht, C., & Vedaldi, A. (2020). Self-labelling via simultaneous clustering and representation learning. In *International conference on learning representations (ICLR)*.
- Aubreville, M., Stathonikos, N., Bertram, C. A., Klopffleisch, R., Ter Hoeve, N., Ciompi, F., ... others (2023). Mitosis domain generalization in histopathology images—the midog challenge. *Medical Image Analysis*, 84, 102699.
- Azizi, S., Culp, L., Freyberg, J., Mustafa, B., Baur, S., Kornblith, S., ... others (2023). Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7(6), 756–779.

- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS one*, 10(7), e0130140.
- Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd international conference on learning representations (iclr)*. Retrieved from <https://arxiv.org/abs/1409.0473>
- Bansal, Y., Nakkiran, P., & Barak, B. (2021). Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, 34, 225–236.
- Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6541–6549).
- Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., ... others (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22), 2199–2210.
- Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., ... Olah, C. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. (<https://transformer-circuits.pub/2023/monosemantic-features/index.html>)
- Bussmann, B., Leask, P., & Nanda, N. (2024). Batchtopk sparse autoencoders. *arXiv preprint arXiv:2412.06410*.
- Byra, M., Galperin, M., Ojeda-Fournier, H., Olson, L., O’Boyle, M., Comstock, C., & Andre, M. (2019). Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. *Medical physics*, 46(2), 746–755.
- Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., ... Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8), 1301–1309.
- Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (eccv)* (pp. 132–149).

- Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33, 9912–9924.
- Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9650–9660).
- Chattopadhyay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (wacv)* (pp. 839–847).
- Chefer, H., Gur, S., & Wolf, L. (2021a, October). Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF international conference on computer vision (iccv)* (p. 397-406).
- Chefer, H., Gur, S., & Wolf, L. (2021b). Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 782–791).
- Chen, C.-F. R., Fan, Q., & Panda, R. (2021). CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In *International conference on computer vision (iccv)*.
- Chen, J., Song, L., Wainwright, M., & Jordan, M. (2018). Learning to explain: An information-theoretic perspective on model interpretation. In *International conference on machine learning* (pp. 883–892).
- Chen, R. J., Chen, C., Li, Y., Chen, T. Y., Trister, A. D., Krishnan, R. G., & Mahmood, F. (2022). Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16144–16155).
- Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F., Jaume, G., Chen, B., . . . others (2024). Towards a general-purpose foundation model for computational pathology. *Nature Medicine*.
- Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607).

- Chen, T., Kornblith, S., Swersky, K., Norouzi, M., & Hinton, G. E. (2020). Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33, 22243–22255.
- Chen, X., Duan, Y., Houthoofd, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29.
- Chen, X., Fan, H., Girshick, R., & He, K. (2020). Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.
- Chen, X., & He, K. (2021). Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15750–15758).
- Chen, X., Xie, S., & He, K. (2021). An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9640–9649).
- Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., ... Liu, J. (2020). Uniter: Universal image-text representation learning. In *European conference on computer vision* (pp. 104–120).
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (p. 248-255). doi: 10.1109/CVPR.2009.5206848
- Ding, T., Wagner, S. J., Song, A. H., Chen, R. J., Lu, M. Y., Zhang, A., ... Mahmood, F. (2024). *Multimodal whole slide foundation model for pathology*. Retrieved from <https://arxiv.org/abs/2411.19666>
- Doersch, C., Gupta, A., & Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. In *2015 IEEE international conference on computer vision (ICCV)* (p. 1422-1430). doi: 10.1109/ICCV.2015.167
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International conference on learning representations*. Retrieved from <https://openreview.net/forum?id=YicbFdNTTy>

- Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., ... others (2022). Toy models of superposition. *arXiv preprint arXiv:2209.10652*.
- Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Kain, A. M., ... Schiratti, J.-B. (2023). Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*. Retrieved from <https://www.medrxiv.org/content/early/2023/07/26/2023.07.21.23292757> doi: 10.1101/2023.07.21.23292757
- Filiot, A., Jacob, P., Kain, A. M., & Saillard, C. (2024). *Phikon-v2, a large and public feature extractor for biomarker prediction*. Retrieved from <https://arxiv.org/abs/2409.09173>
- Fillioux, L., Boyd, J., Vakalopoulou, M., Cournède, P.-H., & Christodoulidis, S. (2023). Structured state space models for multiple instance learning in digital pathology. In *International conference on medical image computing and computer-assisted intervention* (pp. 594–604).
- Fu, R., Hu, Q., Dong, X., Guo, Y., Gao, Y., & Li, B. (2020). *Axiom-based grad-cam: Towards accurate visualization and explanation of cnns*.
- Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., ... Wu, J. (2025). Scaling and evaluating sparse autoencoders. In *The thirteenth international conference on learning representations*. Retrieved from <https://openreview.net/forum?id=tcsZt9ZNKD>
- Gheflati, B., & Rivaz, H. (2022). Vision transformers for classification of breast ultrasound images. In *2022 44th annual international conference of the ieee engineering in medicine & biology society (embc)* (pp. 480–483).
- Gidaris, S., Singh, P., & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. In *International conference on learning representations*. Retrieved from <https://openreview.net/forum?id=Slv4N2l0->
- Glorot, X., Bordes, A., & Bengio, Y. (2011, 11–13 Apr). Deep sparse rectifier neural networks. In G. Gordon, D. Dunson, & M. Dudík (Eds.), *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (Vol. 15, pp. 315–323). Fort Lauderdale, FL, USA: PMLR. Retrieved from <https://proceedings.mlr.press/v15/glorot11a.html>
- Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., ... others (2020).

- Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33, 21271–21284.
- Gu, A., & Dao, T. (2024). Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*. Retrieved from <https://openreview.net/forum?id=tEYskw1VY2>
- Gu, A., Dao, T., Ermon, S., Rudra, A., & Ré, C. (2020). Hippo: Recurrent memory with optimal polynomial projections. *Advances in neural information processing systems*, 33, 1474–1487.
- Gu, A., Goel, K., & Ré, C. (2022). Efficiently modeling long sequences with structured state spaces. In *The international conference on learning representations (ICLR)*.
- Gu, A., Johnson, I., Goel, K., Saab, K., Dao, T., Rudra, A., & Ré, C. (2021). Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34, 572–585.
- Gurnee, W., Nanda, N., Pauly, M., Harvey, K., Troitskii, D., & Bertsimas, D. (2023). Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research*.
- He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 16000–16009).
- He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 9729–9738).
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).
- Hernandez, A., Dangovski, R., & Lu, P. Y. (2022). Model stitching: Looking for functional similarity between representations. In *I can’t believe it’s not better workshop: Understanding deep learning through empirical falsification*. Retrieved from <https://openreview.net/forum?id=Qr5IPOpnLqu>
- Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., ... Lerchner, A. (2017).

- beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*.
- Hosseini, M. S., Bejnordi, B. E., Trinh, V. Q.-H., Chan, L., Hasan, D., Li, X., ... Plataniotis, K. N. (2024). Computational pathology: A survey review and the way forward. *Journal of Pathology Informatics*, 15, 100357. Retrieved from <https://www.sciencedirect.com/science/article/pii/S2153353923001712> doi: <https://doi.org/10.1016/j.jpi.2023.100357>
- Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., ... others (2019). Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1314–1324).
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.
- Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).
- Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J., & Zou, J. (2023). A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9), 2307–2316.
- Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P. K., ... Shapiro, L. (2023). Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36, 37995–38017.
- Ilse, M., Tomczak, J., & Welling, M. (2018, 10–15 Jul). Attention-based deep multiple instance learning. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (Vol. 80, pp. 2127–2136). PMLR. Retrieved from <https://proceedings.mlr.press/v80/ilse18a.html>
- Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., ... Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning* (pp. 4904–4916).

- Jiang, P.-T., Zhang, C.-B., Hou, Q., Cheng, M.-M., & Wei, Y. (2021). Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30, 5875–5888.
- Kim, J., Kim, H. J., Kim, C., Lee, J. H., Kim, K. W., Park, Y. M., ... Kim, W. H. (2021). Weakly-supervised deep learning for ultrasound diagnosis of breast cancer. *Scientific reports*, 11(1), 24382.
- Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Similarity of neural network representations revisited. In *International conference on machine learning* (pp. 3519–3529).
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.
- Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., ... others (2020). The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7), 1956–1981.
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (2002). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.
- Lee, H., Battle, A., Raina, R., & Ng, A. (2006). Efficient sparse coding algorithms. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems* (Vol. 19). MIT Press. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2006/file/2d71b2ae158c7c5912cc0bbde2bb9d95-Paper.pdf
- Li, B., Li, Y., & Eliceiri, K. W. (2021). Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14318–14328).
- Li, Y., Yosinski, J., Clune, J., Lipson, H., & Hopcroft, J. (2016). Convergent learning: Do different neural networks learn the same representations? In *International conference on learning representation (iclr '16)*.
- Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Eccv 2014* (pp. 740–755). Retrieved from https://doi.org/10.1007/978-3-319-10602-1_48 doi: 10.1007/978-3-319-10602-1_48

- Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., ... Liu, Y. (2024). VMamba: Visual state space model. In *The thirty-eighth annual conference on neural information processing systems*. Retrieved from <https://openreview.net/forum?id=ZgtLQQR1K7>
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022).
- Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf
- Lu, M. Y., Chen, B., Williamson, D. F., Chen, R. J., Liang, I., Ding, T., ... others (2024). A visual-language foundation model for computational pathology. *Nature Medicine*, 30, 863–874.
- Lu, M. Y., Williamson, D. F., Chen, T. Y., Chen, R. J., Barbieri, M., & Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6), 555–570.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Mairal, J., Bach, F., Ponce, J., & Sapiro, G. (2009). Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning* (pp. 689–696).
- Makhzani, A., & Frey, B. (2013). K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*.
- Molin, J., Fjeld, M., Mello-Thoms, C., & Lundström, C. (2015). Slide navigation patterns among pathologists with long experience of digital review. *Histopathology*, 67(2), 185–192.
- Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining non-linear classification decisions with deep taylor decomposition. *Pattern recognition*, 65, 211–222.
- Morcos, A., Raghu, M., & Bengio, S. (2018). Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems*,

- Muhammad, M. B., & Yeasin, M. (2020). Eigen-cam: Class activation map using principal components. In *2020 international joint conference on neural networks (ijcnn)* (pp. 1–7).
- Nasiri-Sarvi, A., Hosseini, M. S., & Rivaz, H. (2024). Vision mamba for classification of breast ultrasound images. In *Deep breast workshop on ai and imaging for diagnostic and treatment challenges in breast care* (pp. 148–158).
- Nasiri-Sarvi, A., Rivaz, H., & Hosseini, M. S. (2025). Sparc: Concept-aligned sparse autoencoders for cross-model and cross-modal interpretability. *arXiv preprint arXiv:2507.06265*.
- Nasiri-Sarvi, A., Trinh, V. Q.-H., Rivaz, H., & Hosseini, M. S. (2024). Vim4path: Self-supervised vision mamba for histopathology images. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 6894–6903).
- Noroozi, M., & Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In *Computer vision – eccv 2016* (pp. 69–84). Springer. doi: 10.1007/978-3-319-46466-4_5
- Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23), 3311–3325. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0042698997001697> doi: [https://doi.org/10.1016/S0042-6989\(97\)00169-7](https://doi.org/10.1016/S0042-6989(97)00169-7)
- Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., ... Bojanowski, P. (2024). DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*. Retrieved from <https://openreview.net/forum?id=a68SUt6zFt> (Featured Certification)
- Petsiuk, V., Das, A., & Saenko, K. (2018). Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*.
- Quail, D. F., & Joyce, J. A. (2013). Microenvironmental regulation of tumor progression and metastasis. *Nature medicine*, 19(11), 1423–1437.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

- Raghu, M., Gilmer, J., Yosinski, J., & Sohl-Dickstein, J. (2017). Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In I. Guyon et al. (Eds.), *Advances in neural information processing systems 30* (pp. 6076–6085). Curran Associates, Inc. Retrieved from <http://papers.nips.cc/paper/7188-svcca-singular-vector-canonical-correlation-analysis-for-deep-learning-dynamics-and-interpretability.pdf>
- Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., & Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks? In A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems*. Retrieved from <https://openreview.net/forum?id=R-616EWWKF5>
- Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramár, J., ... Nanda, N. (2024). Improving dictionary learning with gated sparse autoencoders. *arXiv preprint arXiv:2404.16014*.
- Ramaswamy, H. G., et al. (2020). Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 983–991).
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). ” why should i trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144).
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 32).
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4510–4520).
- Saygin Seyfioglu, M., Ikezogwo, W. O., Ghezloo, F., Krishna, R., & Shapiro, L. (2023). Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. *arXiv e-prints*, arXiv–2312.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of*

- the ieee international conference on computer vision* (pp. 618–626).
- Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al. (2021). Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34, 2136–2147.
- Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In *International conference on machine learning* (pp. 3145–3153).
- Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.
- Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd international conference on learning representations, ICLR 2015, san diego, ca, usa, may 7-9, 2015, conference track proceedings*. Retrieved from <http://arxiv.org/abs/1409.1556>
- Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.
- Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.
- Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning* (pp. 3319–3328).
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1–9).
- Tan, H., & Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 conference on empirical methods in natural language processing*.
- Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105–6114).
- Tang, W., Huang, S., Zhang, X., Zhou, F., Zhang, Y., & Liu, B. (2023). Multiple instance learning framework with masked hard instance mining for whole slide image classification. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 4078–4087).

- Thasarakthan, H., Forsyth, J., Fel, T., Kowal, M., & Derpanis, K. G. (2025). Universal sparse autoencoders: Interpretable cross-model concept alignment. In *Forty-second international conference on machine learning*. Retrieved from <https://openreview.net/forum?id=UoaxRN88oR>
- Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *International conference on machine learning* (pp. 10347–10357).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
- Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Liu, S., ... others (2023). Virchow: A million-slide digital pathology foundation model. *arXiv preprint arXiv:2309.07778*.
- Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., ... Hu, X. (2020). Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 24–25).
- Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., ... Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 568–578).
- Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., ... Han, X. (2022). Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81, 102559.
- Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., ... others (2024). A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035), 970–978.
- Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., & Cao, Y. (2021). Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

- Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., & Zhang, L. (2021). Cvt: Introducing convolutions to vision transformers. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 22–31).
- Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., ... others (2025). A vision–language foundation model for precision oncology. *Nature*, 1–10.
- Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., ... Hu, H. (2022). Simmim: A simple framework for masked image modeling. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 9653–9663).
- Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., ... Poon, H. (2024). A whole-slide foundation model for digital pathology from real-world data. *Nature*.
- Yang, S., Wang, Y., & Chen, H. (2024). Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In *International conference on medical image computing and computer-assisted intervention* (pp. 296–306).
- Yap, M. H., Pons, G., Marti, J., Ganau, S., Sentis, M., Zwigelaar, R., ... Marti, R. (2017). Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE journal of biomedical and health informatics*, 22(4), 1218–1226.
- Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.-H., ... Yan, S. (2021). Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 558–567).
- Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning* (pp. 12310–12320).
- Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer vision–eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014, proceedings, part i 13* (pp. 818–833).
- Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S. E., & Zheng, Y. (2022). Dtfdmil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 18802–18812).

- Zhang, R., Isola, P., & Efros, A. A. (2016). Colorful image colorization. In *Computer vision – eccv 2016* (Vol. 9907, pp. 649–666). Springer. doi: 10.1007/978-3-319-46487-9_40
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2921–2929).
- Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). Scene parsing through ade20k dataset. In *Proceedings of the ieee conference on computer vision and pattern recognition*.
- Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., & Kong, T. (2022). ibot: Image bert pre-training with online tokenizer. *International Conference on Learning Representations (ICLR)*.
- Zhu, C., Chai, X., Xiao, Y., Liu, X., Zhang, R., Yang, Z., & Wang, Z. (2024). Swin-net: A swin-transformer-based network combing with multi-scale features for segmentation of breast tumor ultrasound images. *Diagnostics*, 14(3). Retrieved from <https://www.mdpi.com/2075-4418/14/3/269> doi: 10.3390/diagnostics14030269
- Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., & Wang, X. (2024). Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Forty-first international conference on machine learning*.
- Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., ... others (2024). Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*.