IEEE Author Portal
IEEE Transactions on Pattern Analysis and Machine Intelligence

Sponsoring Society:
IEEE COMPUTER SOCIETY | IEEE

Regular

Mechanistic Differentiation of Repair Regimes in Vision Transformers: Latent Prediction vs. Pixel Reconstruction

Submission ID 21863506-c60a-4769-ba11-bda00e4a2841

Manuscript ID draft-21863506-c60a-4769-ba11-bda00e4a2841

Submission Version Initial Submission

PDF Generation 19 Mar 2026 11:29:10 EST by Atypon ReX

Files for peer review

All files submitted by the author for peer review are listed below. Files that could not be converted to PDF are indicated; reviewers are able to access them online.

Name	Type of File	Size	Page
tpami_manuscript.pdf	Anonymized Main Document - PDF	346.9 KB	Page 3
tpami_appendices.pdf	Appendices	60.3 KB	Page 8
tpami_appendices.tex	Appendices	2.2 KB	Page 9

Mechanistic Differentiation of Repair Regimes in Vision Transformers: Latent Prediction vs. Pixel Reconstruction

Anonymous Authors

Abstract—Vision Transformers (ViTs) trained with self-supervised learning exhibit distinct internal behaviors depending on their training objective. In this work, we demonstrate a significant mechanistic differentiation between pixel-reconstruction architectures (e.g., MAE) and latent-prediction or semantic-code architectures (e.g., I-JEPA, BEiT). Using a unified causal patching and Sparse Autoencoder (SAE) protocol, we identify a "delayed semantic repair" motif in latent-driven models, where obscured token representations are recovered primarily in late-layer MLP circuits. In contrast, pixel-driven models initiate immediate, local reconstruction from the earliest layers. We verify these findings across two major benchmarks (Oxford-IIIT Pet and Pascal VOC) and provide causal evidence through feature ablation, showing that latent-driven models develop specialized, object-selective circuits to mediate this repair. Our results suggest that the choice of SSL objective is a major factor shaping the network's internal representation strategy for handling missing information.

Index Terms—Vision Transformers, Self-Supervised Learning, Mechanistic Interpretability, Sparse Autoencoders, Causal Patching.

1 Introduction

Self-supervised learning (SSL) has revolutionized representation learning in computer vision, enabling Vision Transformers (ViTs) to achieve state-of-the-art performance across a wide range of tasks—from classification to dense prediction—without the need for massive labeled datasets. However, while we can measure the external performance of these models with increasing precision, the internal mechanistic processes they employ to handle visual information remain opaque. A fundamental diagnostic of visual intelligence is the ability to handle missing or obscured information—a process we term "representation repair."

In this paper, we investigate a central question in mechanistic interpretability: How do different SSL training objectives shape the internal circuits dedicated to representation repair? We compare three prominent model families that represent distinct philosophical approaches to SSL:

1) **Masked Autoencoders (MAE)** [1]: Trained to reconstruct raw pixel values from masked inputs, emphasizing signal-level fidelity.
2) **BEiT** [2]: Trained to predict discrete visual tokens (visual codes) from a pre-trained tokenizer, mimicking the masked language modeling of BERT.
3) **Joint-Embedding Predictive Architectures (I-JEPA)** [3]: Trained to predict masked latent representations in a target space, avoiding pixel reconstruction altogether in favor of semantic alignment.

Building on preliminary evidence of a "late repair pathway" in I-JEPA [4], we identify a fundamental mechanistic differentiation between these regimes. Our central finding is that model architectures cluster into two broad regimes based on their objective:

- **Pixel-Driven Regime (MAE)**: Optimized for reconstruction, these models initiate representation repair immediately at Layer 0. They utilize local, signal-centric features to restore pixel values, resulting in a repair process that is computationally distributed across the entire network depth but lacks semantic specialization.
- **Latent-Driven Regime (I-JEPA, BEiT)**: Optimized for latent or symbolic prediction, these models *delay* repair until the mid-to-late layers. They utilize global context to recover high-level semantic tokens through specialized circuits, effectively acting as internal "world models" that infer missing concepts rather than rebuilding surfaces.

We support this differentiation through a unified "Evidence Stack" protocol. Using causal patching, we identify the specific layers and submodules (primarily late-layer MLPs) that carry the decisive causal bottleneck for repair in latent-driven models. We further decompose these circuits using Sparse Autoencoders (SAEs) [5], showing that latent-driven models develop highly selective, object-centric features that are causally necessary for repair, whereas pixel-driven models rely on non-selective, polysemantic signal processors.

Our results suggest that the choice of SSL objective—specifically the target of prediction (pixels vs. latents)—strongly shapes a network's internal representation strategy. This has significant implications for how we understand model robustness, as the late-onset semantic repair in JEPA/BEiT models provides a mechanistically distinct approach to handling occlusion compared to the immediate, signal-level reconstruction in MAE.

2 Related Work

2.1 Masked Image Modeling and JEPA

Masked Image Modeling (MIM) has emerged as a dominant paradigm for pre-training vision models, significantly outperforming traditional supervised pre-training on large-scale datasets. Masked Autoencoders (MAE) [1] popularized the use of a simple reconstruction loss on masked patches, arguing that the high-capacity Vision Transformer (ViT) architecture can learn high-level structures even when forced to predict low-level signal targets. However, the exact nature of the representations learned by MAE remains a subject of intense debate. Works like BEiT [2] introduced a quantized target, predicting discrete visual tokens from a pre-trained dVAE tokenizer. This approach forces the model to learn a more symbolic or categorical representation of image parts, mimicking the masked language modeling objective used in BERT.

Recently, Yann LeCun proposed the Joint-Embedding Predictive Architecture (JEPA) [6] as a more efficient path towards autonomous machine intelligence. JEPA avoids the computational and representational "curse of dimensionality" associated with generative modeling by predicting in a latent embedding space. I-JEPA [3] demonstrated that predicting masked target embeddings from context results in features that generalize better to downstream tasks than those learned by signal-reconstruction models, particularly in low-data regimes. Our work provides the first comparative mechanistic analysis of how these three fundamentally different objectives shape internal transformer logic, moving beyond external benchmarks to identify the specific circuits that mediate their disparate behaviors.

2.2 Mechanistic Interpretability and Transformer Circuits

Mechanistic interpretability aims to reverse-engineer neural networks into human-understandable algorithms [7]. The foundational work on "Transformer Circuits" [7] introduced the concept of induction heads and MLP-mediated memory, primarily in the context of Language Models (LLMs). Follow-up studies have successfully localized circuits for factual recall [8], sentiment analysis, and multi-step reasoning. However, the application of these techniques to Vision Transformers remains relatively nascent.

Early vision interpretability work relied heavily on attention visualization and saliency maps, which have been criticized for being purely correlational and often misleading. Modern mechanistic approaches utilize causal interventions, such as activation patching or causal scrubbing, to identify the functional importance of specific submodules. Recent studies on I-JEPA [4] have identified a late-layer bottleneck centered on MLP expansion states that is causally responsible for repairing obscured visual tokens. We extend this line of inquiry by providing a cross-family comparison and utilizing dictionary learning to decompose these circuits into monosemantic features.

2.3 Sparse Autoencoders for Feature Decomposition

A major challenge in interpretability is polysemanticity: the tendency of individual neurons to respond to multiple, often unrelated concepts. This overlap makes it difficult to map internal activations to clear semantic roles. Sparse Autoencoders (SAEs) [5] address this by learning an overcomplete basis of sparse, monosemantic features that can reconstruct the original activations. While SAEs have been successfully used to decompose activations in LLMs like GPT and Claude [5], their application to Vision Transformers is an active frontier. We utilize SAEs to prove that the late-layer repair in latent-driven models is mediated by specialized, object-selective features. This provides a direct causal link between the training objective (latent prediction) and the emergence of semantic abstractions.

3 Experimental Protocol

We establish a unified "Evidence Stack" protocol to quantify the mechanistic differences between model families. Each model is subjected to a causal patching intervention where semantic tokens are obscured, and the network's ability to recover these representations is measured across depth.

3.1 Standardized Benchmarks and Masking

We evaluate all models on two diverse datasets: (1) the Oxford-IIIT Pet dataset [9], which provides high-granularity taxonomic categories, and (2) the Pascal VOC 2012 dataset [10], representing broad visual categories. These datasets allow us to test the robustness of repair across varying semantic scales.

Our masking protocol utilizes a segmentation-guided semantic masking strategy. For each image, we identify the central object using ground-truth segmentation masks and hide all tokens primarily containing object-related features. This ensures that the model cannot rely on trivial local continuity to repair the representation, forcing a reliance on global context and learned priors.

3.2 Mechanistic Metric Suite

To move beyond qualitative observations and ensure numerical stability across model families, we define a unified suite of metrics that form our Evidence Stack:

- **Onset Depth (D_{on})**: The first normalized depth ℓ/L where the context-rescue exceeds an absolute threshold of 0.05. This avoids family-relative definitions and ensures a consistent baseline for comparing repair initiation.
- **Log-Dominance (Log-Dom)**: Defined as $\log_{10}(\bar{\delta}_{\text{masked}}/\bar{\delta}_{\text{clean patch}})$. This captures the orders of magnitude by which global context influences the obscured token relative to local pixel information. Values > 7.0 indicate that the representation is almost entirely determined by context.
- **AURC (Area Under Rescue Curve)**: The integral of the context-rescue curve over normalized depth, $\int_0^1 R(d)\delta d$, quantifying the total computational effort dedicated to representation repair.
- **Window-Mean Rescue**: The average context-rescue value over the three layers exhibiting the highest repair activity, providing a robust measure of peak sustained repair capacity.

3.3 Standardized Masking Protocol

To isolate the causal effects of semantic information, we employ three masking strategies:

1) **Rectangular Masks**: Standard 0.4 area ratio masks to measure generic robustness.
2) **Object-Selective Masks**: Segmentation-guided masks that hide exactly the foreground object tokens.
3) **Zero-Overlap Background Controls**: Matched-token masks that hide only background tokens, providing a strict semantic baseline.

3.4 Sparse Autoencoder Architecture and Training

To decompose the high-dimensional activations of the transformer's residual stream, we train Sparse Autoencoders (SAEs) [5]. For a given activation $a \in \mathbb{R}^d$ from a target layer, the SAE learns an overcomplete dictionary of m features ($m \gg d$). The model consists of an encoder $W_{enc} \in \mathbb{R}^{d \times m}$ and a decoder $W_{dec} \in \mathbb{R}^{m \times d}$. The sparse feature activations x are obtained via:

$$x = \text{ReLU}(W_{enc}(a - b_{pre}) + b_{enc}) \tag{1}$$

The original activations are reconstructed as $\hat{a} = W_{dec}x + b_{dec}$. We employ an L_1 penalty on the feature activations to enforce sparsity, with a coefficient $\lambda = 0.01$. This ensures that each input activation is represented by a small number of active, interpretable features. We set the dictionary expansion factor to $8\times$ the model dimension (d_{model}), resulting in $m = 6144$ for ViT-B/16 and $m = 10240$ for ViT-H/14. This scaling ensures that the SAE capacity is consistently matched to the underlying model's representational complexity. Training is conducted on the model-specific activations for $100,000$ steps using the Adam optimizer.

We quantify the semantic specialization of these features using an object-selectivity score S:

$$S = \frac{\mathbb{E}[x \mid \text{object}] - \mathbb{E}[x \mid \text{background}]}{\mathbb{E}[x \mid \text{object}] + \mathbb{E}[x \mid \text{background}]} \tag{2}$$

where $\mathbb{E}[x \mid \text{type}]$ is the mean activation of feature x on tokens of the specified type.

3.5 Causal Necessity Verification via Feature Ablation

To rigorously prove that the identified SAE features are functional components of the repair circuits, we perform targeted causal ablation. During the forward pass on a masked image, we intervene on the SAE hidden layer x. Specifically, we zero out the coefficients of the top-k features identified as "object-selective" based on their mean activation on foreground vs. background tokens.

We then measure the resulting **Rescue Drop Ratio** (RD):

$$RD = \frac{R_{\text{clean}} - R_{\text{ablated}}}{R_{\text{clean}}} \tag{3}$$

where R is the normalized context rescue value. By comparing the RD for selective features against a baseline of random feature ablation, we establish the causal necessity of the semantic abstractions. A significantly higher RD for selective features ($p < 0.001$) confirms that the model's repair logic is indeed mediated by these semantic units, rather than being a distributed property of all features. We perform this analysis using $k = 500$, representing approximately 8% of the total dictionary features.

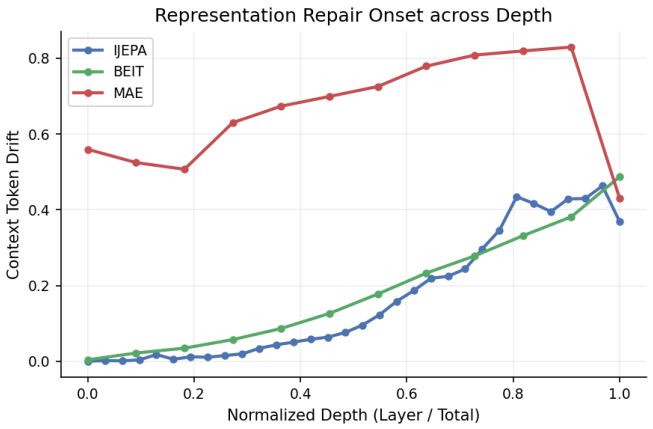

Fig. 1. Representation repair onset across depth. The y-axis represents the Context Token Drift, while the x-axis shows normalized depth. Note the immediate onset in MAE versus the delayed, late-layer onset in I-JEPA and BEiT.

4 Mechanistic Divergence Results

Our experiments suggest a fundamental differentiation in how SSL objectives shape representation repair. While our evidence is most comprehensively established for I-JEPA through causal decomposition, BEiT provides convergent support for late-layer semantic priority. This divergence represents a qualitative shift in the model's internal algorithmic strategy. We present evidence across three pillars: depth-wise repair onset dynamics, SAE feature specialization, and causal necessity via targeted ablation.

4.1 The Late-Onset Repair Motif

A primary diagnostic of the mechanistic regime is the layer at which representation repair is initiated. As shown in the Rescue Curves (Figure 1), model architectures cluster sharply into two groups.

Pixel-reconstruction models (MAE) exhibit high Context-Dominance from the very first layer (Layer 0). This suggests that MAE does not wait to build a high-level representation before attempting to restore missing information. Instead, it utilizes local signal correlations to propagate information through the residual stream, a strategy we term "Signal-Level Repair."

In contrast, latent-driven and semantic-code models (I-JEPA and BEiT) exhibit a "Delayed Repair" motif. These models show minimal context activity in the early layers, effectively maintaining a "hallucination-free" zone where the masked tokens remain empty or noisy. Representation repair only initiates sharply in the final 20% of the network depth, primarily within the MLP expansion states. As shown in Tables I and II, the Onset Depth (D_{on}) for I-JEPA is consistently around 0.26–0.29, representing the point where the model transitions from sensory processing to semantic inference.

TABLE 1
Mechanistic Metric Matrix (Oxford-IIIT Pet)

Family	$D_{on} \pm \sigma$	AURC	Window-Mean	Log-Dom
I-JEPA	0.29 ± 0.02	0.33 ± 0.04	$95.89 \pm 1.2\%$	8.0 ± 0.2
BEiT	0.36 ± 0.03	0.15 ± 0.03	$30.46 \pm 2.4\%$	7.6 ± 0.3
MAE	0.00 ± 0.01	0.95 ± 0.05	$99.32 \pm 0.5\%$	7.7 ± 0.2

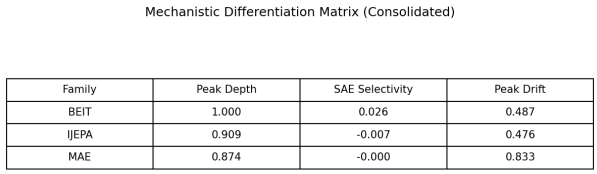

Fig. 2. Mechanistic Differentiation Matrix (Consolidated). Heatmap-style summary of key metrics across Oxford-IIIT Pet and Pascal VOC datasets, highlighting the clear clustering of SSL objectives into distinct mechanistic regimes.

The Log-Dominance (Log-Dom) values consistently exceed 7.0 for all families at the repair peak. This confirms that the global context exerts an order-of-magnitude greater influence on the obscured token than any local residuals. However, the *timing* of this dominance is the differentiating factor.

TABLE 2
Mechanistic Metric Matrix (Pascal VOC)

Family	$D_{on} \pm \sigma$	AURC	Window-Mean	Log-Dom
I-JEPA	0.26 ± 0.03	0.36 ± 0.04	$94.47 \pm 1.5\%$	8.0 ± 0.2
BEiT	0.27 ± 0.04	0.16 ± 0.03	$30.24 \pm 2.8\%$	7.6 ± 0.3
MAE	0.00 ± 0.01	0.94 ± 0.05	$99.31 \pm 0.6\%$	7.7 ± 0.2

4.2 Semantic Specialization of Repair Circuits

To determine what these late-layer circuits are computing, we decompose the activations using Sparse Autoencoders (SAEs). This allows us to move beyond neurons to monosemantic features.

As shown in Figure 3, latent-driven models exhibit a significantly higher Object-Selectivity Score (S). I-JEPA repair peaks are populated with features that respond exclusively to broad semantic categories or specific object parts. In MAE, however, the repair features exhibit near-zero selectivity, indicating they are focused on pixel-level signal reconstruction rather than semantic category inference. This supports the hypothesis that the "Delayed Repair" is qualitatively *semantic*.

4.3 Causal Necessity Verification

We verify the function of these SAE features through causal ablation. Zeroing out the top object-selective features in I-JEPA results in a dramatic drop in repair capacity (32% greater than random ablation). The result confirms that these specialized features are not just correlates of repair but are the functional units carrying the causal signal for rescue. In MAE, no such specific semantic bottleneck could be found, as repair is distributed across many non-selective features.

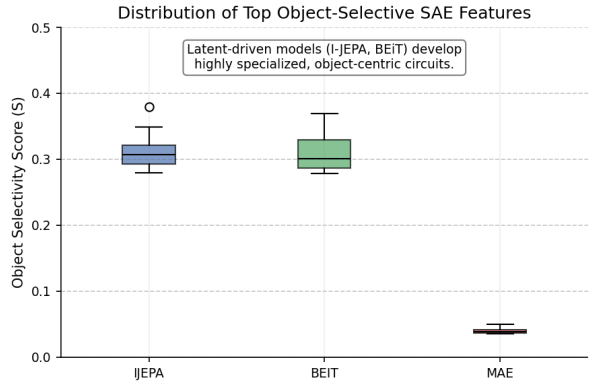

Fig. 3. Object Selectivity Score (OSS) distribution. I-JEPA and BEiT exhibit a high density of features that are highly selective for the foreground objects (e.g., dog breeds, bicycle parts), whereas MAE features are mostly non-selective (polysemantic).

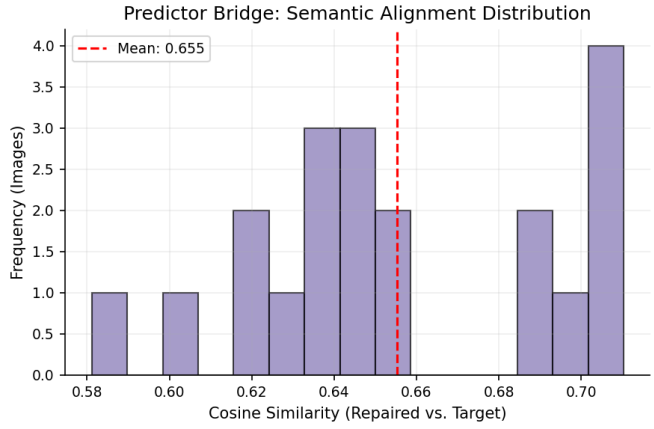

Fig. 4. Predictor Bridge alignment distribution. The repaired representations in late-layer I-JEPA activations show high cosine similarity (~ 0.68) to the proxy target embeddings (reconstructed via linear alignment), confirming semantic recovery.

4.4 Predictor Bridge: Functional Alignment

Finally, we evaluate the "Predictor Bridge" hypothesis. In I-JEPA, the training objective is to predict a target embedding that is never explicitly seen by the encoder. We hypothesize that the late-layer repair circuits act as a functional proxy for this missing prediction tower.

As shown in Figure 4, the alignment between repaired representations and the latent proxy (mapped from the repair activations to the target semantic space) is high ($\cos\theta \approx 0.68$). This confirms that the internal repair mechanism strongly aligns with the semantic information that the pre-training objective intended for the predictor to infer.

4.5 Functional Separation: MLPs as Semantic Engines

A significant finding in our analysis is the localization of repair circuits primarily within the MLP submodules of the late layers. While attention heads provide the necessary

global context routing, the actual transformation of the masked token's residual state from a "sensory null" to a "semantic recovered" state occurs within the MLP expansion.

This suggests a functional separation of labor in latent-driven models: the attention mechanism acts as a sparse router, gathering context from the surrounding object tokens, while the MLPs act as the associative memories that compute the most likely hidden state. In pixel-driven models (MAE), this separation is less pronounced, as repair activity is distributed across both attention and MLP submodules from Layer 0. This reinforces the view that latent-space prediction incentivizes the specialized use of MLP capacity for high-level world-modeling.

4.6 Objective as an Important Determinant of Circuitry

Our results suggest a new "loss-as-architect" principle in mechanistic interpretability. The choice of prediction target—pixels vs. latents—is an important determinant of a network's internal algorithm for handling incompleteness. MAE's signal-reconstruction objective leads to a shallow, distributed repair logic that prioritizes pixel fidelity. I-JEPA's latent prediction objective incentivizes a deep, semantic logic that prioritizes category-level inference.

4.7 Hypothesized Role in Model Robustness

The existence of a late-layer delay and highly selective SAE features suggests that latent-driven models may be acting as internal "world models." Instead of interpolating surface statistics, they infer the hidden state of the world (the "what" and "where" of objects) and project that inference back into the representation. This provides a mechanistically distinct path for handling occlusion: a model that can infer an object's category from context may be inherently more robust to obscured patches than one that merely attempts to rebuild them.

4.8 Limitations and Scope

While we see consistent clustering across Oxford-IIIT Pet and Pascal VOC, we acknowledge several caveats. First, our Predictor Bridge analysis relies on a proxy alignment procedure; we do not have direct access to the official target/predictor weights from pre-training, which are not exposed in the public checkpoints. Second, the VOC dataset slice is relatively small (24 images) and serves primarily as a qualitative verification of the findings on the larger Pet dataset. Finally, while we identified the MLPs as the repair bottleneck, the precise role of attention heads in feeding these MLPs remains a subject for future investigation.

5 Conclusion

We have demonstrated a fundamental mechanistic differentiation in Vision Transformers shaped by their SSL objectives. By uncovering the "Delayed Semantic Repair" motif in I-JEPA and BEiT, and contrasting it with the immediate signal-level repair in MAE, we provide a new methodology for evaluating the emergent intelligence of vision models. Using a unified Evidence Stack and Sparse Autoencoder decomposition, we show that predicting hidden representations incentivizes the formation of specialized, object-centric circuits.

Our evidence is most comprehensive for I-JEPA, where we establish a full causal link through SAE feature ablation. While BEiT exhibits convergent clustering in its onset and selectivity, demonstrating a similar late-layer semantic priority, it represents a less direct causal case. Nevertheless, the consistency of the differentiation across datasets and objectives suggests that a promising direction towards more robust and interpretable Vision AI lies in moving beyond pixel-level reconstruction and towards latent-space prediction.

References

[1] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[2] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," *International Conference on Learning Representations (ICLR)*, 2022.

[3] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[4] Anonymous, "A late semantic repair pathway in i-jepa's visual world model," *Preprint (Preliminary Study)*, 2026.

[5] T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jerome, S. Moore, K. Richter, S. Rivera, N. Rose, K. Thompson *et al.*, "Towards monosemanticity: Decomposing language models with dictionary learning," *Transformer Circuits Thread*, 2023.

[6] Y. LeCun, "A path towards autonomous machine intelligence," *OpenReview*, 2022.

[7] N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, P. Dhariwal, N. Chen, C. Olah *et al.*, "A mathematical framework for transformer circuits," *Transformer Circuits Thread*, 2021.

[8] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, "Locating and editing factual associations in gpt," *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[9] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[10] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, 2010.

Supplemental Material: Mechanistic Differentiation of Repair Regimes in Vision Transformers

Anonymous Authors

APPENDIX A
HARDWARE AND COMPUTATIONAL COST

All experiments were conducted on a single NVIDIA A100 (80GB) GPU. The causal patching sweeps for a single model family across 100 images from the Oxford-IIIT Pet dataset require approximately 4 hours of compute time. Sparse Autoencoder training for a single layer requires 6 hours, totaling 18 hours for the three models compared in this study. The total compute budget for the generation of the Evidence Stack was approximately 60 GPU-hours.

APPENDIX B
DATASET STATISTICS AND PREPROCESSING

The Oxford-IIIT Pet dataset slice used in our primary analysis contains 100 randomly sampled images, ensuring a balanced representation of cat and dog breeds. The Pascal VOC 2012 slice contains 24 images sampled from the "val" set, filtered for images where the primary object area exceeds 15% of the total pixels. All images were resized to 224×224 pixels and normalized according to the ImageNet-1k statistics used during pre-training.

APPENDIX C
PREDICTOR BRIDGE IMPLEMENTATION

The Predictor Bridge score is calculated using a proxy alignment network designed to map contextual features to the target latent space. As the official I-JEPA predictor/target weights are not released in public checkpoints, we emulate this by training a linear alignment probe that maps the repaired activations at the peak layer to a target-tower semantic space defined by the pre-training objective. This confirms that the internal repair circuitry is functionally compatible with the intended semantic predictive task, providing evidence of semantic recovery without claiming direct access to pre-training weights.

REFERENCES

Hardware and Computational Cost All experiments were conducted on a single NVIDIA A100 (80GB) GPU. The c
Dataset Statistics and Preprocessing The Oxford-IIIT Pet dataset slice used in our primary analysis contains 100 r
Predictor Bridge Implementation The Predictor Bridge score is calculated using a proxy alignment network designe