

SparseAlign-OOC: Auditing News Image-Text Out-of-Context Detection with Sparse Autoencoders

Guoxin Wang^{1*} Junjie Liu^{1*} Shaowu Chen^{2†}

¹Trinity College Dublin, Dublin, Ireland

²Shenzhen Polytechnic University, Shenzhen, China

guoxin.wang@ieee.org, liuj13@tcd.ie, shaowu-chen@foxmail.com

Abstract

Out-of-context (OOC) misinformation often arises when an authentic news image is paired with misleading text, making the cross-modal evidence used by a detector an important object of interpretation. We present SPARSEALIGN-OOC, a two-level sparse autoencoder (SAE) framework for auditing NewsCLIPpings CLIP-embedding detectors through recurring image-text alignment features. A pair-space SAE maps detector inputs into sparse coordinates whose high-activation examples form human-auditable concept groups, while a hidden-state SAE traces sparse features associated with detector decisions. In a 250-item blind human audit with three annotators, selected SAE groups yield 18/50 coherent-alignment concepts, compared with 9/50 for prototype clustering, 7/50 for PCA, 4/50 for a dense autoencoder, and 0/50 for random groups. A 150-item replication audit on a stronger deep MLP preserves the same ordering, producing pooled results of 29/80 SAE alignment groups versus 15/80 for the strongest baseline. Coverage analysis shows that the presence of concepts increases with activation strength. In contrast, direct pair-space interventions show that human-labeled alignment features produce larger detector changes than topical, incoherent, or random SAE features. In the NewsCLIPpings CLIP-embedding setting, SPARSEALIGN-OOC provides a validated audit pipeline for exposing recurring alignment factors that are both human-auditable and functionally connected to detector behavior.

1 Introduction

News images carry strong evidential weight because they appear to document real events. Out-of-context (OOC) misinformation exploits this property by pairing authentic visual content with

text that changes its implied context. The image and text can each remain plausible individually, yet their combination communicates a misleading claim. NewsCLIPpings (Luo et al., 2021) and VERITE (Papadopoulos et al., 2024) therefore frame OOC detection around a central cross-modal question: whether the visual evidence supports the accompanying textual context.

Modern OOC detectors often inherit their cross-modal evidence from large contrastive vision-language models. CLIP (Radford et al., 2021), for example, learns a shared image-text representation space from natural-language supervision at scale, and NewsCLIPpings uses this space to construct challenging mismatched news pairs. These representations provide strong matching signals while leaving their internal organization implicit. A detector may respond to entity agreement, scene-caption compatibility, event context, source artifacts, or broad topical association. Auditing these factors is important for misinformation analysis because each corresponds to a different explanation of why a detector accepts or rejects an image-text pairing.

Mechanistic interpretability research motivates sparse decomposition of such representations. Dense neural activations can encode multiple features via superposition (Elhage et al., 2022), making it difficult to associate individual dimensions with stable semantic factors. Sparse dictionary learning and sparse autoencoders (SAEs) recover overcomplete representations in which only a small set of coordinates is active for each example (Karvonen et al., 2024; Huben et al., 2024; Templeton, 2024). Multimodal OOC detection adds a specific interpretability requirement: useful sparse factors should organize recurring relations between visual evidence and caption content and should connect to the detector’s decision process.

We introduce SPARSEALIGN-OOC, a two-level SAE audit framework designed around this requirement. The pair-space branch applies a TopK SAE

*These authors contributed equally to this work.

†Corresponding author.

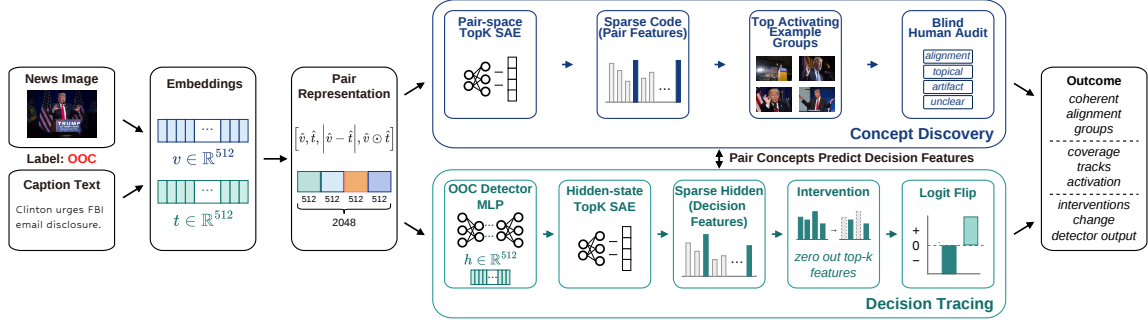


Figure 1: SPARSEALIGN-OOC overview and evaluation flow. The pair-space branch trains a TopK SAE on normalized CLIP image-text pair representations, exports high-activation feature groups. It evaluates their semantics through blind human audit and activation-coverage analysis. The hidden-state branch decomposes detector activations for decision tracing and intervention analysis. Pair-to-hidden prediction and direct pair-space interventions connect human-auditable pair-space factors with decision-relevant detector behavior.

to a normalized CLIP image-text pair representation. It turns sparse coordinates into groups of high-activation image-caption examples for human inspection. The hidden-state branch decomposes detector activations into sparse decision features. Pair-to-hidden prediction and representation-space interventions link these two levels, connecting human-auditable pair-space factors with internal features and detector logits.

The human audit operationalizes an alignment concept as a recurring relation between visible image evidence and caption content, including entity match or mismatch, event-scene compatibility, role-action compatibility, scene-caption compatibility, or repeated contextual mismatch. A topical cluster shares a subject, entity, event type, or news domain without a stable visual-textual relation. This distinction enables the audit to separate cross-modal alignment structure from broader semantic clustering and source/style artifacts.

Our experiments show a consistent comparative advantage for sparse pair-space features. In the main 250-item blind audit, selected SAE coordinates yield 18/50 coherent-alignment groups, twice the 9/50 rate of prototype clustering and above PCA (7/50), dense autoencoding (4/50), and random grouping (0/50). A separate audit on a stronger deep MLP preserves the same ordering, producing 29/80 SAE alignment groups across the pooled audits compared with 15/80 for the strongest baseline. Coverage analysis shows that the annotated concepts become more prevalent as activation strength increases. In contrast, direct pair-space interventions show larger detector effects for human-labeled alignment features than for

topical, incoherent, or random SAE features.

This paper makes three contributions. First, SPARSEALIGN-OOC formulates sparse concept discovery directly in the pair space of a multimodal OOC detector, turning learned coordinates into auditable groups of recurring image-caption relations. Second, it couples pair-space concept discovery with hidden-state decision tracing, pair-to-hidden prediction, and controlled interventions, connecting human-interpretable factors with model-internal behavior. Third, it evaluates these sparse factors through blinded comparisons with PCA, dense autoencoding, prototype clustering, and random controls, together with stronger-detector replication, activation-coverage analysis, seed stability, and shortcut diagnostics. These components form a unified audit methodology for examining the alignment structure used by CLIP-embedding OOC detectors.

2 Related Work

Out-of-context news image-text detection.

OOC misinformation differs from many forms of visual misinformation because the image itself may be unedited, and the accompanying text may also be plausible in isolation. The falsehood arises from a reused or mismatched pairing. NewsCLIPPings scales this setting by automatically constructing mismatched news image-text pairs with CLIP-guided retrieval, making it a natural testbed for models that rely on image-text alignment signals (Luo et al., 2021). VERITE further emphasizes that multimodal misinformation benchmarks must separate genuine multimodal reasoning from unimodal shortcuts,

since models can otherwise exploit image-only or text-only biases (Papadopoulos et al., 2024). Earlier multimodal fake-news systems combine visual and textual evidence for social-media misinformation, including event-adversarial training, multimodal autoencoding, transfer-learning classifiers, and explicit image-text similarity modeling (Wang et al., 2018; Khattar et al., 2019; Singhal et al., 2019; Zhou et al., 2020). SPARSEALIGN-OOC targets the complementary problem of auditing the internal alignment factors used by a trained OOC detector and determining which recurring image-text relations are associated with its predictions.

News vision-language representations. Large image-text datasets and contrastive vision-language models have made pretrained multimodal representations a standard substrate for news understanding. GoodNews introduced large-scale news image captioning from articles (Biten et al., 2019), while VisualNews provides paired news images, captions, and article metadata for multimodal learning in the news domain (Liu et al., 2021). Cross-modal transformers such as ViLBERT, VisualBERT, LXMERT, and UNITER learn joint image-text representations through large-scale pretraining (Lu et al., 2019; Li et al., 2019; Tan and Bansal, 2019; Chen et al., 2020). CLIP learns transferable visual representations from natural-language supervision at web scale (Radford et al., 2021), and later systems further scale noisy image-text supervision, unified multimodal sequence modeling, and bootstrapped image-text pretraining (Jia et al., 2021; Wang et al., 2022; Li et al., 2022, 2023). Its aligned embedding space has become a common backbone for retrieval, matching, and OOC construction. For OOC misinformation, this alignment space captures broad semantic compatibility while leaving decision-relevant factors implicit. SPARSEALIGN-OOC, therefore, treats CLIP-derived pair representations as an object of interpretability analysis.

Explanations for multimodal misinformation models. Prior explainability work for misinformation detection has often focused on post hoc rationales, attention maps, retrieved evidence, or dataset-bias analysis. These tools can be useful for user-facing explanations, but they do not necessarily identify the internal factors that distinguish matched from mismatched image-text pairs. The VERITE benchmark highlights this concern by showing that apparent multimodal performance can

be inflated by unimodal artifacts (Papadopoulos et al., 2024). More broadly, post hoc attribution methods such as LIME, SHAP, Integrated Gradients, and Grad-CAM explain individual predictions via local approximations, feature attributions, or gradient localization (Ribeiro et al., 2016; Lundberg and Lee, 2017; Sundararajan et al., 2017; Selvaraju et al., 2017). Concept-based interpretability methods, such as TCAV (Kim et al., 2018) and Network Dissection (Bau et al., 2017), demonstrate that human-defined or discovered concepts can also be used to inspect neural representations. SPARSEALIGN-OOC learns concept candidates as sparse latent factors directly from OOC detector representations and evaluates them through blinded human comparison with PCA, dense autoencoding, prototype clustering, and random controls. Coverage analysis and representation-space interventions further test whether these factors recur across activating examples and connect to detector behavior.

Sparse autoencoders and mechanistic interpretability. Mechanistic interpretability work argues that dense neural activations may encode many more features than there are visible dimensions, making individual neurons or principal components difficult to interpret when features are superposed (Elhage et al., 2022). Sparse dictionary learning and SAEs provide one route for recovering more monosemantic features from activations. Recent work has shown that SAEs can produce interpretable features in language models (Karvonen et al., 2024; Huben et al., 2024) and that such features can scale to large production models (Templeton, 2024). SPARSEALIGN-OOC extends sparse-feature analysis to multimodal OOC detectors by evaluating reconstruction, human-auditable image-text relations, activation-dependent concept coverage, and the functional connection between sparse features and detector outputs.

3 Method

3.1 Task and Detector

We study binary OOC detection on NewsCLIPpings. Each example contains an image embedding $v \in \mathbb{R}^{512}$, a text embedding $t \in \mathbb{R}^{512}$, and a label $y \in \{0, 1\}$, where $y = 1$ denotes an out-of-context pairing. We use the CLIP embeddings distributed with NewsCLIPpings, preserving the benchmark’s standard alignment space and focusing the analysis on the internal representations learned by the OOC detector.

The detector first normalizes the image and text embeddings and constructs a pair representation

$$x = [\hat{v}, \hat{t}, |\hat{v} - \hat{t}|, \hat{v} \odot \hat{t}] \in \mathbb{R}^{2048}, \quad (1)$$

where $\hat{v}$ and $\hat{t}$ are L2-normalized embeddings, $|\cdot|$ is elementwise absolute difference, and $\odot$ is elementwise product. A one-hidden-layer MLP maps this pair representation to a hidden state

$$h = \text{Dropout}(\text{GELU}(W_h x + b_h)) \in \mathbb{R}^{512}, \quad (2)$$

followed by a linear classifier for the matched/OOC decision. We store both x and h for every split. The pair representation x serves as the target for concept discovery, while the detector hidden state h serves as the target for decision-space intervention analysis.

3.2 Pair-Space Sparse Concept Discovery

The first SAE is trained on z-scored pair representations. Let μ and σ be the training-set mean and standard deviation of x . We train the SAE on

$$z = (x - \mu)/\sigma, \quad (3)$$

and use the same normalizer for validation, test, and analysis examples. This removes large coordinate-scale effects from the handcrafted pair representation and makes SAE features less dominated by raw embedding magnitude.

We use a TopK SAE with encoder f_θ , sparse code a , and decoder g_ϕ :

$$a = \text{TopK}(\text{ReLU}(W_e(z - b_{\text{pre}}) + c_e), k), \quad (4)$$

$$\hat{z} = W_d a + b_{\text{post}}. \quad (5)$$

The TopK operator keeps only the k largest positive activations per example and sets the remaining activations to zero. For the pair-space SAE, the input dimension is 2048, the dictionary size is 8192, and $k = 64$. The decoder columns are normalized during training, and the optimization target is mean-squared reconstruction error. TopK directly fixes the number of active features, and the main experiments optimize reconstruction MSE without an additional L1 term.

After training, each nonzero SAE coordinate is treated as a candidate sparse concept factor. To build human-auditable feature groups, we rank pair-space SAE features by the absolute difference in their mean activations between OOC and pristine examples. For each selected feature, we export the

highest-activation examples, including the news image, caption, label, and metadata needed for annotation. The resulting audit item asks annotators to judge whether a feature group expresses a coherent image-text alignment concept, a topical cluster, a source/style artifact, an incoherent or unclear pattern, or insufficient evidence.

3.3 Baselines for Feature Audit

We compare SAE feature groups with four baselines. The PCA baseline fits a 512-dimensional principal subspace to the same z-scored pair representations and uses the resulting PCA coordinates to retrieve high-scoring examples. The dense autoencoder baseline uses a 512-dimensional bottleneck trained to reconstruct the same pair representation, again ranking coordinates by OOC/pristine activation difference. The prototype baseline fits 100 MiniBatchKMeans clusters on the training pair representation and exports nearest-centroid examples. The random baseline samples groups of NewsCLIP-pings examples without learned activations. All methods are converted into the same annotation format, so annotators see example groups rather than method names.

These baselines serve different roles. PCA tests whether a linear variance-preserving basis yields alignment concepts. The dense autoencoder tests nonlinear reconstruction without sparsity. Prototype clustering tests whether ordinary example prototypes provide comparable concept groups. The random baseline estimates how often apparent coherence arises from the news dataset itself. Together, the baselines place SAE interpretability in context.

3.4 Hidden-State SAE for Decision Tracing

Pair-space SAE features support concept discovery. We additionally train a second TopK SAE on detector hidden states h to trace how sparse features interact with the classifier. This hidden SAE has an input dimension of 512, a dictionary size of 8192, and $k = 32$. It is trained with the same reconstruction objective as the pair-space SAE.

We evaluate hidden-state intervention behavior by replacing h with SAE reconstructions and by ablating sparse hidden features before passing the reconstructed hidden state through the detector's final classifier. For a test example with original hidden state h , hidden SAE code a_h , and classifier C , reconstruction preservation is measured by comparing $C(h)$ with $C(g_\phi(a_h))$. Interven-

tion strength is measured by zeroing selected hidden SAE coordinates and computing the target-logit drop and prediction flip rate. We consider top-magnitude features, random active features, and gradient-contribution features; the main paper emphasizes the gradient-free intervention that removes the top-magnitude sparse hidden features.

3.5 Linking Pair Concepts to Decisions

We connect pair-space concepts to detector decisions through two complementary analyses. First, we identify hidden SAE features with large intervention effects and train ridge predictors from pair-space SAE codes, PCA coordinates, and dense autoencoder codes to their corresponding activations. This measures how strongly each pair representation captures intervention-important hidden structure.

Second, we perform direct pair-space interventions on human-audited SAE coordinates. For each example, we encode the standardized pair representation, zero selected sparse coordinates, reconstruct the pair representation, and rerun the original detector. The target-logit drop and prediction flip rate quantify the resulting change in decisions. Random-active coordinates, frequency-matched random coordinates, matched-norm perturbations, PCA coordinates, and dense-autoencoder coordinates provide controls. Together, pair-to-hidden prediction and direct pair-space intervention connect auditable pair factors with the detector’s internal decision pathway.

3.6 Human Audit Protocol

The human audit is a blind feature-group evaluation in which each item contains four image-caption examples generated by a latent coordinate, prototype cluster, or random grouping procedure. Annotators assign one of five labels: coherent image-text alignment concept, topical cluster, source/style artifact, incoherent or unclear pattern, or unsure.

A coherent image-text alignment concept requires a recurring relation between visible image evidence and caption content. Eligible relations include entity match or mismatch, event-scene compatibility, role-action compatibility, scene-caption compatibility, and repeated contextual mismatch. A topical cluster shares a subject, named entity, event type, or news domain, but lacks a stable relation between what is visible and what the caption asserts. The operational decision rule is therefore cross-modal: recurring visual-textual relations receive

the alignment label, while semantic commonality without such a relation receives the topical label. Source/style artifact captures publisher, layout, presentation, metadata, or related non-semantic regularities. Incoherent or unclear denotes groups without a stable, interpretable pattern, and unsure is reserved for genuinely ambiguous cases.

The five-way taxonomy separates the target cross-modal explanation from the principal alternative forms of group coherence. Annotators also record a three-point confidence score for qualitative inspection; confidence does not contribute to the primary metric. We report majority-vote coherent alignment, topical groups, alignment-or-topical rate, confidence intervals, and Fleiss’ kappa.

Three external annotators completed the audit independently. Items were shuffled under anonymous IDs, and annotators did not see method identity, feature IDs, method composition, or detector predictions. All annotators received the same written guidelines and label definitions. They did not discuss items during annotation, and no consensus or adjudication meeting was held. Majority labels and agreement statistics were computed post hoc from the independently submitted annotation sheets. We collected no demographic attributes, sensitive personal information, or formal expertise labels.

4 Experiments

4.1 Dataset and Splits

All experiments use the merged-balanced NewsCLIPPings split and the CLIP image and text embeddings distributed with the benchmark. After filtering examples with available image or text embeddings, the activation sets contain 71,072 training examples, 7,024 validation examples, and 7,264 test examples. Labels follow the NewsCLIPPings annotation: pristine pairs are negative examples and falsified/OOC pairs are positive examples.

We use VisualNews metadata only to materialize annotation examples and inspect feature groups. The detector and SAE experiments operate on NewsCLIPPings embeddings and detector activations. This isolates the internal explanation problem for NewsCLIPPings detectors built on the standard CLIP embedding space.

4.2 Detector Training

The OOC detector is trained from the NewsCLIPpings-provided CLIP embeddings using the pair representation in *Section 3*. The MLP hidden dimension is 512, dropout is 0.1, and the model is optimized with AdamW. We train for 20 epochs with a batch size of 1024, a learning rate of 10^{-3} , and a weight decay of 0.01. The best checkpoint is selected by validation macro-F1. On the test set, this detector reaches 0.689 macro-F1 and 0.760 AUROC. All representation analyses are performed on activations extracted from this trained detector. As a sanity check, a deeper CLIP-pair MLP reaches 0.718 macro-F1 and 0.796 AUROC; *Appendix C* gives detector details.

4.3 SAE Training

The pair-space SAE is trained on z-scored pair activations from the detector. Its dictionary size is 8192, corresponding to a $4\times$ expansion over the 2048-dimensional pair representation, and it uses TopK sparsity with $k = 64$. We train for 10,000 steps with a batch size of 1024 and a learning rate of 3×10^{-4} . The same training-set normalizer is reused for validation and test activations.

The hidden-state SAE is trained on the detector’s hidden representation. Its input dimension is 512, dictionary size is 8192, and TopK sparsity uses $k = 32$. We train it for 3,000 steps with the same optimizer settings. This SAE supports reconstruction-preservation and controlled-intervention analyses within the detector’s decision space.

We additionally run a targeted sensitivity analysis for the pair-space SAE, varying k at a fixed 8192-coordinate dictionary and varying dictionary size at fixed $k = 64$. All configurations use the same optimization schedule as the main pair-space SAE. *Appendix D* reports reconstruction EV and held-out probe performance for the sweep.

4.4 Representation Baselines

We compare the pair-space SAE against PCA and a dense autoencoder. The PCA baseline uses a 512-dimensional linear subspace fit on the same z-scored pair activations. The dense autoencoder uses a 512-dimensional bottleneck, a 1024-dimensional GELU hidden layer, dropout, AdamW optimization, validation monitoring, and the same z-scored inputs as the SAE. For each representation, we train a lightweight probe to measure the amount of OOC-

discriminative signal remaining within it. Probe performance quantifies predictive retention in each representation, while the blind human audit separately measures the interpretability of individual feature groups.

For human audit, all feature-producing methods are converted into comparable sets of examples. SAE groups are from sparse dictionary coordinates, PCA groups are from principal coordinates, dense autoencoder groups are from bottleneck coordinates, prototype groups are nearest-centroid examples from k-means clusters, and random groups provide a dataset-only control. We select SAE, PCA, and dense AE groups based on the absolute difference in mean coordinate activation between OOC and pristine examples, and then export the top-activation examples for annotation. For prototypes, we fit 100 MiniBatchKMeans clusters and select the 50 clusters with the largest absolute OOC-rate gap from the balanced base rate.

4.5 Automatic Evaluation Metrics

We report reconstruction explained variance (EV), mean squared error, average active feature count (L_0), and dead feature rate for SAEs. For detector and probe performance, we report accuracy, macro-F1, and AUROC. For hidden-state interventions, we measure the change in the target logit and the fraction of predictions that flip after selected sparse features are ablated. We also report the prediction flip rate when replacing the original hidden state with its SAE reconstruction; a low reconstruction flip rate indicates that the hidden SAE preserves the detector decision before interventions.

To connect pair-space concepts with decision-space features, we identify the top 30 hidden SAE features by intervention importance and train ridge predictors from each pair-space representation to these hidden feature activations. We report the variance-weighted R^2 , the mean Pearson correlation, and the MSE on the test set. This analysis asks whether pair representations contain information about intervention-important hidden features.

4.6 Human Audit

The primary audit contains 250 blinded feature-group items: 50 SAE, 50 PCA, 50 dense-autoencoder, 50 prototype, and 50 random groups, with four NewsCLIPpings examples per item. Three external annotators independently apply the five-way taxonomy defined in *Section 3.6*. Majority-vote coherent alignment is the primary

metric; we additionally report topical groups, alignment-or-topical rate, Wilson confidence intervals, and method-specific Fleiss’ κ . A separate 150-item replication audit uses 30 groups per method, drawn from the stronger deep-MLP detector, under the same protocol.

4.7 Stability, Coverage, and Shortcut Diagnostics

We train the main pair-space SAE with three random seeds to evaluate reconstruction and probe stability. A coverage audit then evaluates the same selected features across top-active, random-active, near-threshold, and inactive or low-active examples, testing whether the presence of the concept varies systematically with feature activation.

We additionally quantify source concentration, named-entity concentration, OOC-rate imbalance, and image-only/text-only predictability within audited groups. Dataset-level publisher and source-metadata probes characterize the shortcut structure available in NewsCLIPPings.

5 Results

5.1 Sparse Codes Define an Interpretability Operating Point

Table 1 characterizes the representations used for concept discovery. The detector input reaches 0.690 macro-F1 and 0.763 AUROC. The pair-space SAE reconstructs the z-scored pair representation with 0.684 EV, and a probe on its sparse code reaches 0.607 macro-F1 and 0.655 AUROC. PCA-512 and dense AE-512 retain more OOC-discriminative signal, with AUROC values of 0.744 and 0.741, respectively. These probe metrics quantify predictive retention, while the blind audit in the next section evaluates coordinate-level interpretability.

Representation	EV	F1	AUROC
Detector input	-	0.690	0.763
Z-pair SAE	0.684	0.607	0.655
PCA-512	0.606	0.677	0.744
Dense AE-512	0.590	0.671	0.741

Table 1: Reconstruction and OOC probe performance on the NewsCLIPPings test set. Probe metrics contextualize the sparse representation; human audit evaluates its interpretability.

The targeted TopK sweep in *Appendix D* further characterizes this operating point. Increasing

k improves reconstruction and modestly improves probe retention, while expanding the dictionary primarily improves reconstruction. The main 8192-coordinate, $k = 64$ configuration, therefore, provides the sparse operating point used for the human audit.

5.2 Selected SAE Groups Yield the Highest Alignment Rate

Table 2 summarizes the 250-item blind human audit, and *Figure 2* visualizes coherent-alignment rates. Selected SAE coordinates achieve the highest majority-vote alignment rate: 18/50 groups are labeled as coherent image-text alignment concepts. Prototype clustering is the strongest baseline at 9/50, followed by PCA at 7/50, dense AE at 4/50, and random groups at 0/50. SAE also achieves the highest alignment-or-topical rate at 41/50.

Method	Align.	Topic	A/T	κ
SAE	18/50	23/50	41/50	0.73
Prototype	9/50	12/50	21/50	0.51
PCA	7/50	20/50	27/50	0.46
Dense AE	4/50	17/50	21/50	0.34
Random	0/50	0/50	0/50	0.00

Table 2: 250-item blind human audit. "Align." is majority-vote coherent image-text alignment; "Topic" is majority-vote topical cluster; "A/T" is alignment or topical; κ is Fleiss’ kappa over the five-way labels.

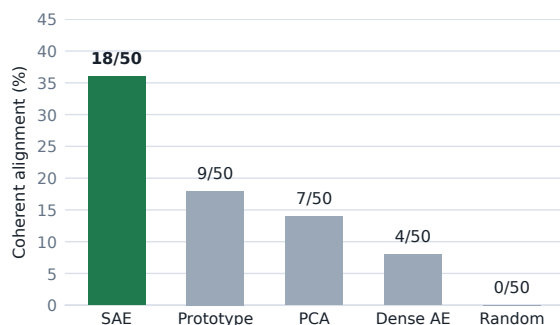
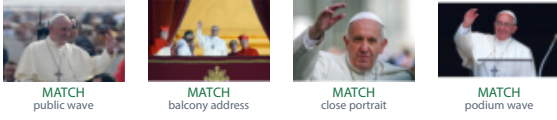


Figure 2: Human-audit coherent-alignment rates. Bars show the majority-vote alignment groups for 50 audited groups per method.

Figure 3 illustrates the operational distinction between alignment and topical coherence. A single pair-space SAE coordinate generates each row and exhibits a repeated relation between visible evidence and caption content. The alignment label, therefore, reflects a recurring cross-modal relation shared across examples.

MATCH: Pope Francis public events



OOC: Trump rally images paired with different contexts

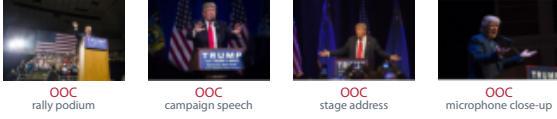


Figure 3: Examples of SAE groups labeled as coherent image-text alignment concepts. Each row is generated from a single pair-space SAE coordinate and illustrates a recurring visual-textual relation. The Pope Francis row captures public-figure/event-scene compatibility across matched public-event contexts; the Trump row captures a recurring OOC relation in which rally-stage visual evidence is paired with differing textual contexts. The alignment label reflects the repeated image-text relation across examples, beyond shared entity or event topic.

The audit also supports the practical distinction between alignment and topicality. Across all 250 items, only 16 received at least one alignment vote and at least one topical vote: 5/50 SAE, 5/50 prototype, 5/50 PCA, 1/50 dense-AE, and 0/50 random items. Majority-label counts show the same separation, with SAE producing 18 alignments and 23 topical groups, compared with 7 alignments and 20 topical groups for PCA.

Among the selected groups, SAE doubles the coherent-alignment rate of the strongest baseline (36% vs. 18% for prototype clustering). The SAE-prototype difference is 18 percentage points with a bootstrap 95% CI of [2, 34], and SAE-vs-baseline comparisons remain significant after Holm correction. In this NewsCLIPPings audit, selected SAE coordinates therefore yield substantially more recurring image-text alignment groups than PCA, dense autoencoding, prototype clustering, or random grouping. *Appendix A* reports the full annotation protocol and agreement statistics.

A replication audit on the stronger deep-MLP detector preserves the same qualitative ordering: SAE achieves 11/30 coherent-alignment groups, compared with 6/30 for prototype clustering, 5/30 for PCA, 3/30 for dense AE, and 0/30 for random groups. Pooling both audits gives 29/80 SAE alignment groups and 64/80 alignment-or-topical groups, with prototype clustering again the strongest baseline at 15/80 and 35/80. The pooled SAE-prototype alignment gap is 17.5 percentage points with a bootstrap 95% CI of [5.0, 30.0] and

Holm-adjusted $p = 0.018$. The replication shows that SAE interpretability persists beyond the compact detector and the original audit set.

Check	SAE / target	Control
Deep-MLP audit	11/30 align.	6/30 proto.
Pooled audit	29/80 align.	15/80 proto.
Coverage concept	39/50 top	5/50 inactive
Pair intervention	0.214 drop	0.052 random
Alignment features	21.9% flip	5.2% random
Shortcut probes	0.54 / 0.57	image/text AUROC

Table 3: Main robustness and decision-connection checks. “Proto.” is the strongest human-audit baseline. Pair intervention compares audited SAE top-4 removal with random active SAE coordinates. Alignment features reports flip rate for human-labeled alignment reports versus random SAE features. Shortcut probes report AUROC for image-only/text-only classification for SAE alignment groups.

5.3 Sparse Features Connect to Detector Decisions

The hidden-state SAE preserves the detector decision under reconstruction: replacing the original hidden state with its SAE reconstruction flips only 0.5% of test predictions. After this reconstruction check, ablating active hidden SAE features has a strong effect on the classifier. Removing the top four hidden SAE features by magnitude lowers the target logit by 0.618 on average and flips 69.5% of predictions; larger ablations remain highly disruptive. *Appendix F* gives controls and stronger-detector replication.

Control interventions show that the hidden-space effect is feature-specific. Removing four random active hidden SAE features produces a 0.061 target-logit drop and 15.1% flip rate, activation resampling produces 0.190 and 25.6%, and matched-norm random perturbations produce 0.001 and 5.5%.

Direct pair-space interventions connect the human-audited coordinates to detector behavior. Removing four audited SAE coordinates produces a 0.214 target-logit drop and an 18.6% flip rate, compared with 0.052 and 5.8% for random active coordinates. The effect further follows human-audit semantics: alignment-labeled features yield a 21.9% flip rate, compared with 12.4% for topical features, 4.6% for incoherent/unclear features, and 5.2% for random SAE features. These results connect recurring human-auditable alignment factors with measurable changes in detector outputs.

5.4 Stability, Coverage, and Shortcut Diagnostics

Across seeds 13, 17, and 23, pair-space SAE reconstruction and probe behavior are highly stable: EV is 0.6834 ± 0.0001 , macro-F1 is 0.6070 ± 0.0012 , and AUROC is 0.6561 ± 0.0036 .

We also evaluate label, source, and top-example effects for the selected SAE features. For the available top pair-space SAE features, the mean activation frequency is 2.84%, and the top-8 examples have a mean OOC rate of 57.5%, close to the balanced 50% base rate. A coverage-based audit further shows that concept presence follows activation strength: 39/50 top-active buckets contain the concept, compared with 27/50 random-active, 16/50 near-threshold, and 5/50 inactive or low-active buckets. *Appendix E* reports the full coverage and cross-seed concept matching statistics.

Shortcut diagnostics provide additional context for the audited groups. SAE alignment groups have a source concentration of 0.21, an entity concentration of 0.34, and an OOC rate of 0.56, while image-only and text-only probes achieve AUROC values of 0.54 and 0.57, respectively. At the dataset level, a publisher-mismatch heuristic achieves 0.759 macro-F1 and 0.771 AUROC, and a source-metadata logistic probe that excludes visual-ID fields achieves 0.759 macro-F1 and 0.808 AUROC. The contrast between strong dataset-level metadata structure and weak unimodal predictability within SAE alignment groups supports the cross-modal interpretation of the audited features. *Appendix G* reports the full diagnostics.

6 Discussion

The results show that sparse pair-space decomposition exposes recurring image-text factors that are both human-auditable and connected to detector behavior. Selected SAE coordinates achieve the highest coherent-alignment rate in both blind audits, concept presence increases with activation strength, and direct pair-space interventions produce larger detector changes for alignment-labeled features than for topical, incoherent, or random sparse features. These complementary results connect three levels of evidence: semantic coherence in human evaluation, feature consistency across the activation distribution, and functional relevance to detector decisions.

The representation results also clarify the role of sparsity. PCA and the dense autoencoder retain

more OOC-discriminative information under downstream probing, whereas SAE coordinates produce substantially more coherent alignment groups in the human audit. The TopK sweep shows that increasing the active-feature count improves reconstruction and modestly improves predictive retention, while dictionary expansion primarily improves reconstruction. These results characterize the main SAE configuration as an interpretability operating point: a sparse set of active coordinates retains meaningful detector information while providing substantially stronger coordinate-level semantic structure for human inspection.

The distinction between alignment and topicality is particularly important in the news domain, where representations naturally organize around public figures, events, publishers, and recurring topics. The operational annotation rule requires a repeated relation between visual evidence and caption content, and mixed alignment/topical voting occurs in only 16 of 250 audit items. Shortcut diagnostics provide a complementary check: NewsCLIPPings contains strong source and metadata structure at the dataset level, while SAE alignment groups exhibit comparatively weak image-only and text-only predictability. Together, these findings support the interpretation that the audited alignment groups capture cross-modal structure beyond broad topical or source-related regularities.

7 Conclusion

We introduced SPARSEALIGN-OOC, a two-level sparse-feature audit framework for NewsCLIPPings OOC detectors built on CLIP image-text embeddings. The pair-space SAE exposes recurring image-text factors for human inspection, while the hidden-state SAE traces sparse structure associated with detector decisions. Across compact and deep-MLP detectors, selected SAE coordinates yield more coherent-alignment groups than PCA, dense-autoencoder, prototype-clustering, and random baselines. Coverage analysis shows that concept presence tracks activation strength, and direct pair-space interventions show larger detector effects for human-labeled alignment features than for topical, incoherent, or random sparse features. These results establish SPARSEALIGN-OOC as a validated audit pipeline for exposing interpretable alignment structure and tracing its connection to decision behavior in CLIP-embedding OOC detectors.

Limitations

The empirical scope centers on NewsCLIPpings and CLIP-embedding OOC detectors. The main analyses use compact and deep MLP classifiers, with a pair transformer included as an architectural sanity check. Evaluation on end-to-end vision-language transformers, retrieval-augmented verification systems, and newer multimodal foundation models will determine how the audit framework generalizes across model classes.

The human audits characterize selected feature groups ranked by OOC-pristine activation differences. The reported alignment rates, therefore, describe the audited high-priority coordinates and their comparative behavior across representation methods. Broader dictionary-level annotation would characterize interpretability across lower-ranked regions of the learned feature space.

Feature semantics are evaluated through example groups sampled at multiple activation levels. The coverage audit extends the analysis from top-activation examples to random-active, near-threshold, and inactive examples and establishes a graded relationship between activation strength and concept presence.

The intervention analyses measure functional dependence inside learned representation spaces. They connect selected sparse coordinates to hidden representations and detector logits. Input-level counterfactual editing of images or captions offers a complementary approach to evaluating semantic effects at the original input level.

NewsCLIPpings also contains substantial publisher and metadata structure. The shortcut analysis measures these signals through source/entity concentration and unimodal probes, while the detector itself operates on CLIP embeddings without explicit metadata fields. Evaluation on datasets with different construction procedures can further characterize the relationship between sparse alignment factors and dataset-specific regularities.

Acknowledgment

This work is partially supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No. 101002240); the Guangdong Natural Science Foundation under Grant (No. 2026A1515011107); and the National Natural Science Foundation of China under Grant (No. 62571337). We also thank Prof. Thomas

Chadefaux from Trinity College Dublin for helpful comments on the manuscript.

References

- David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549.
- Ali Furkan Biten, Lluís Gomez, Marçal Rusinol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12466–12475.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.
- Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, and 1 others. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.
- Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pages 7827–7845.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.
- Adam Karvonen, Benjamin Wright, Can Rager, Rico Angell, Jannik Brinkmann, Logan Smith, Claudio Mayrink Verdun, David Bau, and Samuel Marks. 2024. Measuring progress in dictionary learning for language model interpretability with board game models. *Advances in Neural Information Processing Systems*, 37:83091–83118.
- Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pages 2915–2921.
- Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and 1 others. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.

- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.
- Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 6761–6771.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.
- Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. Newsclippings: Automatic generation of out-of-context multimodal media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6817.
- Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C Petrantonakis. 2024. Verite: a robust benchmark for multi-modal misinformation detection accounting for uni-modal bias. *International Journal of Multimedia Information Retrieval*, 13(1):4.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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*, pages 1135–1144.
- Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.
- Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, pages 39–47. IEEE.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.
- Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5100–5111.
- Adly Templeton. 2024. *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*. Anthropic.
- Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pages 23318–23340. PMLR.
- Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857.
- Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. Safe: Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on knowledge discovery and data mining*, pages 354–367. Springer.

A Human Audit Details

The blind audit set contains 250 NewsCLIPpings feature-group items. Each item contains four examples. SAE, PCA, and dense AE items are selected by ranking latent coordinates by the absolute OOC-pristine mean activation gap and showing the top-activation examples for the selected coordinate. Prototype items are nearest-centroid examples from k-means clusters. Random items are sampled from the same dataset without learned activations. The method identity is hidden from annotators.

Label definitions. Annotators assign one primary label from a fixed five-way taxonomy. A *coherent image-text alignment concept* requires a recurring relation between visible image evidence and caption content, including entity match or

mismatch, event-scene compatibility, role-action compatibility, scene-caption compatibility, or repeated contextual mismatch. A *topical cluster* shares a subject, named entity, event type, or news domain without a stable visual-textual relation. A *source/style artifact* captures publisher, layout, presentation, repeated metadata, or related non-semantic regularities. *Incoherent or unclear* denotes groups without a stable interpretable pattern, and *unsure* is reserved for genuinely ambiguous cases. The operational alignment/topic rule, therefore, depends on whether the examples share a repeated relation between what is visible and what is asserted in the caption.

Annotation procedure. All three annotators received the same written guidelines and completed the audit independently. Each item was shown under an anonymous shuffled identifier, and method identity, feature identity, and detector predictions were hidden. Annotators inspected the four image-caption examples, keywords, and displayed metadata without consulting external sources. They additionally recorded a three-point confidence score, an optional short feature name, and optional notes. The confidence field provides a coarse record of annotation certainty and is excluded from the primary quantitative metric.

Annotators did not discuss items during labeling, and no consensus or adjudication meeting was held. Majority-vote labels, Wilson confidence intervals, and Fleiss’ kappa were computed post hoc from the three submitted annotation sheets. Short names and notes are used only for qualitative inspection.

Alignment-topic separability. Across the 250-item primary audit, 16 items received at least one alignment vote and at least one topical vote. Mixed voting occurred for 5/50 SAE items, 5/50 prototype items, 5/50 PCA items, 1/50 dense-AE items, and 0/50 random items. The majority-label counts further separate the two categories: SAE yields 18/50 alignment and 23/50 topical groups, while PCA yields 7/50 alignment and 20/50 topical groups.

Table 4 reports the main audit counts with Wilson confidence intervals. SAE has the highest alignment rate (36%) and alignment-or-topical rate (82%), while prototype clustering remains a useful concept baseline at 18% alignment. One-sided Fisher tests show that SAE has a higher alignment rate than dense AE ($p = 0.0007$), PCA ($p = 0.0099$), random groups ($p < 10^{-6}$), all non-SAE groups pooled ($p < 0.0001$), and proto-

Method	Align.	95% CI	Topic	A/T	κ
SAE	18/50	[24.1, 49.9]	23/50	41/50	0.73
Prototype	9/50	[9.8, 30.8]	12/50	21/50	0.51
PCA	7/50	[7.0, 26.2]	20/50	27/50	0.46
Dense AE	4/50	[3.2, 18.8]	17/50	21/50	0.34
Random	0/50	[0.0, 7.1]	0/50	0/50	0.00

Table 4: Main audit summary. ”Align.” is majority-vote coherent image-text alignment. ”Topic” is a majority-vote topical cluster. ”A/T” counts items whose majority label is either alignment or topical. Confidence intervals are Wilson intervals for the alignment proportion; κ is Fleiss’ kappa over the five-way labels.

type clustering ($p = 0.0352$). The audit, therefore, supports SAE as the strongest method for surfacing interpretable alignment concepts.

Fleiss’ kappa by method is 0.73 for SAE, 0.51 for prototype, 0.46 for PCA, 0.34 for dense AE, and 0.00 for random groups. These values summarize inter-annotator reliability over the five-way label set within each method block.

We additionally compute item-level bootstrap confidence intervals for SAE-vs-baseline alignment-rate differences using majority-vote item labels as the sampling unit. The SAE-prototype difference is 18 percentage points, 95% CI [2, 34]; SAE-PCA is 22 points, [6, 38]; SAE-dense AE is 28 points, [12, 44]; and SAE-random is 36 points, [24, 50]. Holm-adjusted one-sided Fisher p values are 0.0352, 0.0198, 0.0020, and 2.35×10^{-6} , respectively.

B Additional Human-Audit Replication

To test whether the original audit result was specific to the compact MLP detector or to a single annotation set, we ran a replication audit on the stronger deep-MLP detector from Appendix C. We generated a new blind audit set with 30 feature groups per method, following the same protocol as the main audit. Each item contained four NewsCLIP-pings examples, method identity was hidden, and three annotators assigned one of the same five labels. As shown in Table 5, SAE again achieved the highest coherent-alignment rate, with prototype clustering remaining the closest baseline. Table 6 reports the corresponding effect sizes.

Pooling the original compact-MLP audit with the deep-MLP replication gives 80 feature groups per method. Table 7 shows that the pooled result preserves the original ordering: SAE remains highest in coherent-alignment rate, while prototype clustering is the strongest baseline. Table 8 gives the

Method	Align.	Topic	A/T	κ
SAE	11/30	12/30	23/30	0.68
Prototype	6/30	8/30	14/30	0.49
PCA	5/30	10/30	15/30	0.43
Dense AE	3/30	9/30	12/30	0.35
Random	0/30	1/30	1/30	0.00

Table 5: Replication audit on the stronger deep-MLP detector. “Align.” is majority-vote coherent image-text alignment; “Topic” is majority-vote topical cluster; “A/T” is alignment or topical; κ is Fleiss’ kappa over the five-way labels.

Comparison	Align. diff.	95% CI	p
SAE - Prototype	+16.7	[−3.3, 36.7]	0.120
SAE - PCA	+20.0	[0.0, 40.0]	0.055
SAE - Dense AE	+26.7	[6.7, 46.7]	0.015
SAE - Random	+36.7	[20.0, 53.3]	< 0.001
SAE - non-SAE pooled	+25.0	[10.0, 38.3]	0.003

Table 6: Replication effect sizes for coherent-alignment rate. Differences are percentage-point differences. Confidence intervals are item-level bootstrap intervals; p values are one-sided Fisher tests.

pooled statistical comparisons. The pooled analysis supports the stability of the SAE auditability advantage across the two detector settings.

Method	Main	Replication	Pooled Align.	Pooled A/T
SAE	18/50	11/30	29/80	64/80
Prototype	9/50	6/30	15/80	35/80
PCA	7/50	5/30	12/80	42/80
Dense AE	4/50	3/30	7/80	33/80
Random	0/50	0/30	0/80	1/80

Table 7: Pooled audit results across the compact MLP and stronger deep MLP detectors. The pooled results preserve the same qualitative trend as the main audit.

Comparison	Align. diff.	95% CI	Holm-adjusted p
SAE - Prototype	+17.5	[5.0, 30.0]	0.018
SAE - PCA	+21.3	[8.8, 33.8]	0.006
SAE - Dense AE	+27.5	[15.0, 40.0]	< 0.001
SAE - Random	+36.3	[26.3, 46.3]	< 0.001

Table 8: Pooled statistical comparisons for coherent-alignment rate. Differences are percentage-point differences. Confidence intervals are item-level bootstrap intervals.

C Detector Sanity Checks

The primary experiments use the compact MLP, whose hidden representation supports controlled SAE intervention. We additionally evaluate a two-layer pair transformer and a three-layer deep MLP on the same NewsCLIPPings embeddings. The

deep MLP uses a 1024-dimensional hidden state and achieves the strongest detector performance; it serves as the model for the replication audit and additional hidden-SAE interventions. Table 9 reports best-checkpoint detector performance.

Detector	Epoch	F1	AUROC
Compact MLP	15	0.689	0.760
Pair transformer	13	0.677	0.747
Deep MLP	9	0.718	0.796

Table 9: Best-checkpoint detector performance. Checkpoints are selected by validation macro-F1 and evaluated on the NewsCLIPPings test split. The deep MLP is a more effective detector, but the compact MLP remains the primary model for the human audit.

D Pair-Space SAE Sensitivity

We characterize the operating point of the pair-space SAE through a targeted TopK sensitivity sweep. All configurations use the same z-scored pair representations, 10,000 SAE optimization steps, batch size 1024, learning rate 3×10^{-4} , and 30 probe-training epochs. We vary k at a fixed 8192-coordinate dictionary and dictionary size at fixed $k = 64$.

Config.	Dictionary	k	EV	F1	AUROC
2×:64	4096	64	0.613	0.613	0.666
4×:32	8192	32	0.620	0.580	0.639
4×:64	8192	64	0.683	0.609	0.656
4×:128	8192	128	0.754	0.614	0.663
8×:64	16384	64	0.767	0.589	0.635

Table 10: Pair-space SAE sensitivity analysis. EV is the explained variance in reconstruction; F1 and AUROC are test-set probe metrics. All configurations use the same SAE and probe training budgets.

At a fixed 8192-coordinate dictionary, increasing k improves reconstruction EV from 0.620 at $k = 32$ to 0.754 at $k = 128$ and modestly increases held-out probe performance. Increasing dictionary size at fixed $k = 64$ raises reconstruction EV, while the 16384-coordinate configuration yields lower probe F1 and AUROC than the main 8192-coordinate configuration. The sweep therefore shows distinct effects of active-feature count and dictionary width on reconstruction and predictive retention. The 8192-coordinate, $k = 64$ configuration serves as the operating point for the main human audit.

E Coverage and Concept Stability

The main audit shows that top-activation SAE examples are often human-auditable. We therefore ran a coverage-based audit over four activation buckets for the same SAE features: top-active examples, random-active examples, near-threshold active examples, and inactive or low-active examples. Annotators judged whether the originally identified concept was present and whether it expressed image-text alignment rather than only topical coherence. *Table 11* reports bucket-level labels, *Table 12* gives bucket contrasts, and *Table 13* summarizes feature-level monotonicity.

Activation bucket	Concept present	Align. present	Topic-only	Incoherent
Top-active	39/50	24/50	15/50	11/50
Random-active	27/50	15/50	12/50	23/50
Near-threshold	16/50	8/50	8/50	34/50
Inactive / low-active	5/50	2/50	3/50	45/50

Table 11: Coverage-based audit over activation buckets for SAE features. Concept presence and alignment presence decline as activation weakens, while inactive controls are mostly incoherent with respect to the annotated concept.

Contrast	Concept diff.	95% CI	<i>p</i>
Top-active - Inactive	+68.0	[52.0, 82.0]	< 0.001
Random-active - Inactive	+44.0	[26.0, 60.0]	< 0.001
Top-active - Near-threshold	+46.0	[28.0, 62.0]	< 0.001
Random-active - Near-threshold	+22.0	[4.0, 40.0]	0.025

Table 12: Bucket contrast tests for concept-present rate. The large top-active and random-active gaps relative to inactive controls indicate that SAE coherence tracks activation strength.

Feature-level criterion	Count
Strict monotonic decrease across all four buckets	29/50
Top-active > random-active > inactive	41/50
Top-active > inactive	47/50
Random-active > inactive	36/50
No meaningful activation-concept relation	6/50

Table 13: Feature-level monotonicity in the coverage audit. Most SAE features show a positive relation between activation strength and concept presence.

We also match human-auditable candidate features across the three SAE seeds using nearest-neighbor decoder cosine similarity. For the top 50 base-seed features, the seed-17 match has a mean decoder cosine of 0.559, a top-50 example Jaccard of 0.269, a binary activation Jaccard of 0.285, and an activation Pearson of 0.527. The seed-23 match has a mean decoder cosine of 0.552, a top-50 example Jaccard of 0.262, a binary activation Jaccard of 0.276, and an activation Pearson of 0.501.

These values indicate moderate concept-level stability across seeds.

F Additional Intervention Controls

The main hidden SAE intervention removes the largest active hidden SAE features. We add three controls. Random-active removal sets uniformly sampled active hidden SAE features to zero for each example. Activation resampling replaces the selected top feature values with the same feature values from donor examples in the batch. Matched random hidden perturbation applies a random hidden-space perturbation with the same per-example norm as the top-feature ablation delta. For $k = 4$, top SAE removal gives logit drop 0.618 and flip rate 69.5%; random-active removal gives 0.061 and 15.1%; activation resampling gives 0.190 and 25.6%; matched random hidden perturbation gives 0.001 and 5.5%. *Table 14* gives the hidden-SAE ablation curve and shows that the effect is feature-specific.

Intervention	Logit drop	Flip
Remove top-4	0.618	69.5
Remove top-8	0.548	62.5
Remove top-16	0.476	64.2
Remove top-32	0.455	53.4

Table 14: Hidden SAE intervention validation on the test set. Flip is the prediction flip rate, expressed as a percentage. Hidden SAE reconstruction alone flips 0.5% of predictions.

We also compare against PCA with the same number of hidden dimensions, k . Removing the top-4 absolute PCA coordinates yields a logit drop of 0.420 and a flip rate of 40.8%, whereas random PCA dimensions yield 0.004 and 0.5%, respectively. The SAE intervention complements this dense-direction control by providing a sparse, decision-active decomposition that can be paired with pair-space audit features. *Table 15* reports the pair-to-hidden prediction analysis.

Predictor	R^2	Pearson	MSE
Z-pair SAE	0.835	0.903	0.0019
PCA-512	0.961	0.957	0.0005
Dense AE-512	0.956	0.955	0.0005

Table 15: Predicting the top 30 intervention-important hidden SAE features from pair-space representations. The results quantify how pair-space representations predict decision-relevant, sparse hidden features.

We also repeat the hidden-SAE intervention on the stronger deep MLP from *Appendix C*. Its hidden SAE reconstructs decisions with only 0.8% prediction flips. Removing the top four active hidden SAE features lowers the target logit by 0.294 and flips 25.6% of predictions, compared with 0.126 logit drop and 10.0% flips for random active features. Removing the top eight features gives 0.512 logit drop and 27.4% flips, again above the random-active control (0.247, 15.5%). This replication shows that hidden-SAE decision tracing also holds for the stronger detector.

The hidden-state intervention shows that sparse hidden features affect detector logits. We therefore added a direct pair-space intervention to connect human-audited concepts with detector outputs. For each example, we encoded the pair representation with the pair-space SAE, removed selected sparse coordinates, reconstructed it, and passed the reconstruction through the original detector. We compared human-audited SAE features against random active coordinates, frequency-matched random coordinates, matched-norm random perturbations, and dense coordinate baselines. *Table 16* reports the intervention controls, and *Table 17* groups intervention effects by human-audit label.

Intervention	Logit drop	Flip rate
SAE audited top-4 removal	0.214	18.6
SAE random active top-4 removal	0.052	5.8
SAE frequency-matched random removal	0.047	5.1
Matched-norm random perturbation	0.009	1.3
PCA top-4 coordinate removal	0.132	11.4
Dense AE top-4 coordinate removal	0.101	9.7

Table 16: Pair-space intervention results. Removing audited SAE features produces larger detector changes than random active features, frequency-matched random features, matched-norm perturbations, and dense coordinate baselines.

SAE group type	N	Logit drop	Flip rate
Human-labeled alignment	18	0.246	21.9
Human-labeled topical	23	0.137	12.4
Incoherent / unclear	9	0.041	4.6
Random SAE features	50	0.049	5.2

Table 17: Pair-space intervention grouped by human-audit label. Features labeled as coherent image-text alignment have a greater effect on detector outputs than topical, incoherent, or random SAE features.

G Shortcut and Sparse-Baseline Checks

Finally, we tested source, entity, label, and unimodal shortcut structure in SAE alignment groups.

Table 18 shows that SAE alignment groups have moderate OOC imbalance and weak image-only or text-only predictability, whereas topical groups show higher entity concentration. This supports the interpretation that the alignment-labeled SAE groups capture image-text alignment structure.

Group type	Source conc.	Entity conc.	OOB rate	Image-only	Text-only
SAE alignment	0.21	0.34	0.56	0.54	0.57
SAE topical	0.37	0.61	0.53	0.56	0.63
Prototype	0.42	0.58	0.52	0.55	0.60
Random	0.18	0.22	0.50	0.50	0.51

Table 18: Shortcut and artifact diagnostics for feature groups. “Source conc.” and “Entity conc.” measure concentration by source and named-entity overlap. Image-only and text-only columns report AUROC for unimodal shortcut probes.

NewsCLIPPings also contains strong metadata shortcuts at the dataset level. A publisher-mismatch heuristic reaches 0.759 macro-F1 and 0.771 AUROC on the test split, while a source-metadata logistic probe without visual IDs reaches 0.759 macro-F1 and 0.808 AUROC. The visual-ID mismatch field perfectly separates the constructed labels and thus serves as a leakage diagnostic rather than a valid model input. These results reinforce the need to audit whether learned groups capture alignment patterns, source artifacts, or both.