UNDERSTANDING CANDIDATE-SHARED EVIDENCE IN VISUAL FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

What parts of a CLIP feature matter for a prediction? We find that once the classifier has narrowed an image to a few likely classes, most of the feature can be removed without changing the answer. We take the top-K classes from a frozen classifier and use a sparse autoencoder to split the CLIP feature into concepts. Across 11 datasets, the concepts that all K candidates share are also the most strongly expressed, and they make up roughly two thirds of the feature. We show why the classifier trained on top of CLIP keeps them: its cross-entropy loss over all classes rewards evidence that separates the likely classes from all the others, even when that evidence does nothing to tell the likely classes apart. We then train two small rerankers that only choose among the K candidates, one on the full feature and one with the shared concepts deleted. Both improve three standard few-shot CLIP classifiers, a linear probe and two prompt-learning methods, by 1.3 to 5.9 points at 16 shots, and they match within 0.2 points. The deleted concepts alone reach 13% accuracy among the candidates against 68% for what remains, and the full-feature reranker learns to ignore them on its own, cutting their influence on the score by 74%. Our results show that global recognition relies on evidence the K candidates share, and that local discrimination can discard it and learns to do so on its own.

1 INTRODUCTION

Vision foundation models, such as CLIP (Radford et al., 2021), learn rich, transferable representations, enabling zero-shot recognition for a wide range of downstream tasks as well as efficient adaptation through lightweight classifiers or learned prompts. The information encoded by such models, however, need not be equally relevant to every downstream decision. Classical representation theory relates minimal sufficient representations for a given task to invariance against task-irrelevant variability (Achille & Soatto, 2018); critically, both sufficiency and irrelevance are defined relative to the task. Since a foundation representation commonly remains fixed during downstream adaptation, it is not guaranteed to discard irrelevant information for a particular downstream decision. This broad retention of information is a key source of transferability, but it also raises a fundamental question:

How does a downstream classifier use representation components whose relevance changes with the prediction context?

To examine the internal representations of CLIP, we begin from the phenomenon of *polysemanticity*, whereby a single representation dimension can encode multiple factors (Elhage et al., 2022). Since such mixing in dense CLIP representations makes it difficult to isolate the contribution of individual components, we employ a sparse autoencoder (SAE) (Huben et al., 2024), which expresses the frozen representation in a set of sparsely activated latent atoms. For each input, we obtain a candidate set from the global classifier and measure how frequently and how strongly each sparse component is expressed across the candidate classes. Our analysis reveals that components frequently activated across all candidate classes also exhibit substantially greater activation magnitudes than other components. We refer to this correspondence as *prevalence–magnitude agreement* (Section 2.1), indicating that probable classes systematically share prominent representation components.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

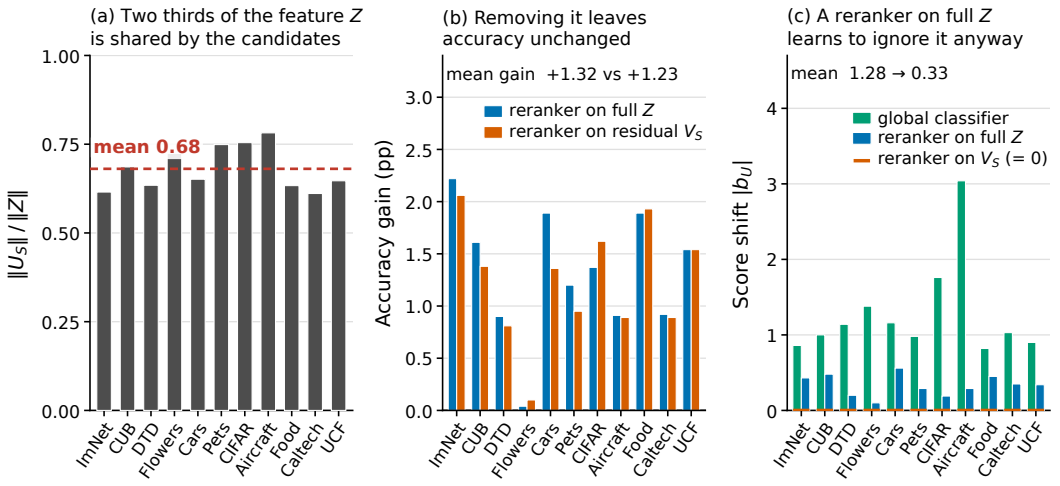


Figure 1: **Evidence shared by the candidate classes is large, dispensable, and already discounted by a candidate-restricted head.** (CLIP ResNet-50, 16-shot linear probe, top-K=15) **(a)** The decoded candidate-shared component U_S accounts for 0.68 (average) of the feature norm $\|Z\|_2$ across eleven datasets. **(b)** A reranker trained on the full feature Z and one trained on the residual $V_S = Z - U_S$ improve the frozen classifier by nearly identical margins (details in Table 1). **(c)** The uncancelled score shift the shared component induces between the true class and its strongest competitor, $|b_U(y, c^\star)|$ (Equation 9), reduces under the full-Z reranker; or zero by construction for the residual reranker. Candidate-restricted training learns to discount the shared evidence.

To explain why candidate-shared components can persist and affect local decisions, we identify a *global–local objective mismatch* (Section 2.2). Conditioned on successful retrieval, global cross-entropy decomposes into within-candidate discrimination and candidate-versus-rest separation. Increasing all candidate logits equally improves the latter but cancels out in the former, allowing global training to reward evidence that provides no discrimination within the candidate set. When such evidence contributes only approximately equally under different class prototypes, the uncancelled deviations can perturb or even reverse candidate-local margins (Section 2.3).

Together, the prevalence–magnitude agreement and the global–local objective mismatch motivate a controlled study (Section 3 and Section 4) separating optimization scope from the necessity of candidate-shared components. We first train a lightweight reranker on the complete frozen representation Z, restricting its objective to the candidate classes retrieved by the global classifier. This removes the candidate-versus-rest term and tests whether objective alignment improves local discrimination. Across eleven recognition datasets and three downstream classifiers, the full-Z reranker consistently improves or preserves the performance of its underlying global classifier. However, because all components of Z remain accessible, any reduced reliance on shared evidence is implicit. We therefore compare it with an architecture-matched reranker operating on V_S, obtained by suppressing identified candidate-shared atoms; the two variants differ only in their input representation. Reranking on V_S performs comparably to reranking on Z, indicating that the selected shared components are largely dispensable for candidate-local discrimination. These results establish candidate conditioning as a useful lens for studying how information relevance changes with prediction scope. Figure 1 summarizes these findings. Our contributions are as follows:

- **Candidate-conditioned representations structure.** We identify and characterize sparse components shared across plausible alternatives, and show strong agreement between their activation prevalence and magnitude across eleven recognition datasets.
- **Objective-level account of global–local mismatch.** We show how globally common evidence can improve candidate-versus-rest separation while canceling from the ideal within-candidate decision, with imperfect commonality capable of perturbing local margins.
- **Mechanistic study**. We compare architecture-matched candidate-restricted classifiers using the full and intervened representations, testing whether selected candidate-shared components are necessary for local discrimination.

2 Problem Setting and Theoretical Basis

2.1 Problem Setting and Candidate-Shared Concepts

Problem setting. Let $(X, Y) \sim \mathcal{P}$ denote an image–label pair with $Y \in \mathcal{C} = \{1, \dots, C\}$. A frozen visual encoder E_{I}, such as the CLIP image encoder, maps X to a representation $Z = E_{\mathrm{I}}(X) \in \mathbb{R}^d$. A global class-prototype learner g_θ assigns each class c a prototype $w_c = g_\theta(c) \in \mathbb{R}^d$; this covers both learned linear classifiers and text prototypes from a prompt learner. The score and predictive probability for class c are

$$s_c(Z) = w_c^\top Z, \qquad p_\theta(Y = c \mid Z) = \frac{\exp s_c(Z)}{\sum_{r \in \mathcal{C}} \exp s_r(Z)}, \tag{1}$$

The global prediction is $\arg\max_{c \in \mathcal{C}} s_c(Z)$. Given Z, the global classifier retrieves the K highest-scoring classes,

$$\mathcal{S}(Z) = \mathrm{TopK}_{c \in \mathcal{C}}\ s_c(Z), \qquad |\mathcal{S}(Z)| = K, \tag{2}$$

which we call the *candidate set* and write as $\mathcal{S}$ when the dependence on Z is clear. We denote successful retrieval of the ground-truth class by $\mathcal{E}_K = \{Y \in \mathcal{S}(Z)\}$.

Sparse concept decomposition. To expose the concepts contributing to Z, we use a sparse autoencoder (SAE), an unsupervised dictionary-learning model that reconstructs Z from a sparse, overcomplete set of latent atoms. Let $A = E_{\mathrm{SAE}}(Z) \in \mathbb{R}^m_{\geq 0}$ be the sparse activation vector, with A_j the activation of atom j, and let $D = [d_1, \dots, d_m] \in \mathbb{R}^{d \times m}$ be the decoder dictionary. Each active atom contributes approximately $A_j d_j$ to the reconstruction of Z and can be read as a reusable visual concept. We use a TopK SAE (Gao et al., 2025), which keeps only the k_{SAE} largest encoder activations; the training procedure is given in Appendix B.

Class prevalence and magnitude. Given a labeled analysis set, we summarize each atom per class by how often it fires and how strongly:

$$\widehat{\pi}_{c,j} = \frac{1}{n_c} \sum_{i : y_i = c} \mathbb{I}[A_j(Z_i) > \delta]\,, \qquad \widehat{\mu}_{c,j} = \frac{\sum_{i : y_i = c} A_j(Z_i)\, \mathbb{I}[A_j(Z_i) > \delta]}{\sum_{i : y_i = c} \mathbb{I}[A_j(Z_i) > \delta]}. \tag{3}$$

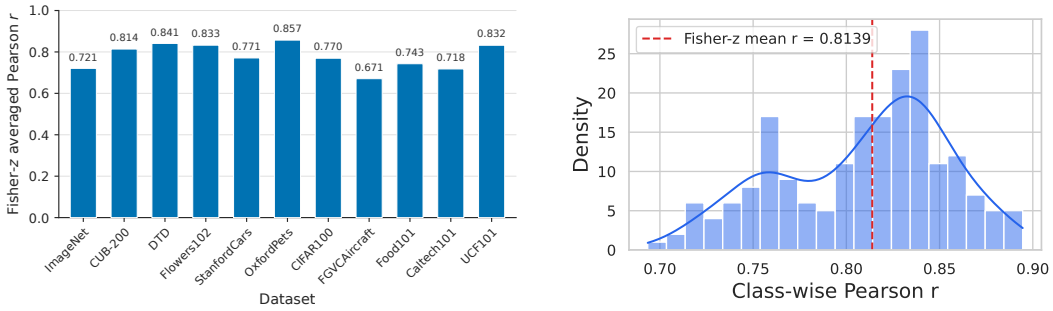
The prevalence $\widehat{\pi}_{c,j}$ drives the selection rule in Section 3; the magnitude $\widehat{\mu}_{c,j}$ is used only in the diagnostic below.

Motivating observation: prevalence–magnitude agreement. Before analyzing the training mismatch, we check whether concepts that fire consistently across probable classes are also expressed strongly. For each target class c we form a probable set $\mathcal{S}_c$ from the $K - 1$ classes most often confused with c, score every atom by how uniformly it is prevalent across $\mathcal{S}_c$ and, separately, by how uniformly it is strong across $\mathcal{S}_c$, and correlate the two scores over atoms (Appendix C). Figure 2 shows strong agreement: the Fisher-z-averaged correlation exceeds 0.7 on ten of eleven datasets, most CUB-200 classes are individually high, and within a single candidate neighborhood the atoms ranked highest by prevalence carry nearly all of the magnitude-based sharedness. Concepts that recur across probable classes are therefore also the concepts expressed most strongly (see Appendix G for other datasets). This suggests that candidate-shared evidence can form a prominent part of the representation, which motivates the two questions below: why global training keeps it, and how its strong but imperfectly shared contribution can interfere with candidate-local discrimination. Agreement alone does not show interference; that mechanism is Proposition 2, tested in Section 4.

2.2 Why Global Training Retains Locally Shared Evidence

Evidence shared by every retrieved candidate gives no relative preference among them, yet global training can still reward it for separating the candidate group from the rest of the label space. For an example with label y, the global classifier minimizes

$$\ell_{\mathcal{C}}(Z, y) = -\log p_\theta(Y = y \mid Z) = -s_y(Z) + \log \sum_{c \in \mathcal{C}} \exp s_c(Z). \tag{4}$$



(a) Dataset-level correlations.

(b) Class-wise correlation distribution on CUB-200.

Figure 2: **Agreement between prevalence- and magnitude-based candidate sharedness.** For each target class, we compare prevalence-based sharedness $Q^P_{\mathcal{S}_c,j}$ with magnitude-based sharedness $Q^M_{\mathcal{S}_c,j}$ over valid SAE atoms (Appendix C).

Fix a candidate set $\mathcal{S} \ni y$ and write $A_\mathcal{S} = \sum_{c\in\mathcal{S}} \exp s_c(Z)$ and $B_\mathcal{S} = \sum_{r\notin\mathcal{S}} \exp s_r(Z)$. The global loss splits exactly into a within-candidate term and a candidate-versus-rest term:

$$\ell_\mathcal{C}(Z,y) = \underbrace{-s_y(Z) + \log A_\mathcal{S}}_{\ell_\mathcal{S}(Z,y)} + \underbrace{\log\left(1 + \frac{B_\mathcal{S}}{A_\mathcal{S}}\right)}_{\text{separation from non-candidates}}. \tag{5}$$

The first term, the candidate-restricted loss $\ell_\mathcal{S}$, ranks classes within $\mathcal{S}$. The second rewards pushing the candidate group away from everything else.

Proposition 1 (Global training rewards candidate-shared evidence)**.** *Fix $\mathcal{S} \subset \mathcal{C}$ with $y \in \mathcal{S}$. Suppose a feature adds a common score α to every candidate and leaves non-candidates unchanged,*

$$s_c^{(\alpha)}(Z) = s_c^{(0)}(Z) + \alpha\, \mathbb{I}[c \in \mathcal{S}]. \tag{6}$$

Then the candidate-restricted loss is invariant to α, while the global loss strictly decreases in α whenever $\mathcal{C} \setminus \mathcal{S}$ is nonempty.

Proof in Appendix D. The common shift cancels from the restricted softmax but shrinks the second term of Equation 5; its derivative equals minus the probability mass the classifier places outside $\mathcal{S}$ (Corollary 1, Appendix D). Proposition 1 exposes the mismatch: evidence that raises all candidate scores together is invisible to the within-candidate decision but still lowers the global loss, so a globally trained classifier keeps it. It is kept for a legitimate reason, locating the right region of the label space, and should not be read as spurious. Its role simply changes with scope: useful for coarse separation over $\mathcal{C}$, non-discriminative for the fine decision within $\mathcal{S}$. An objective normalized over $\mathcal{S}$ alone has no candidate-versus-rest term and so removes this incentive, which motivates the first reranking strategy in Section 3. We note that nothing here says shared evidence is harmful. Harm requires the candidate-specific scores it induces to be approximately, rather than exactly, equal, which is the case studied next.

2.3 Dominance Can Amplify Imperfect Local Cancellation

Evidence shared at the concept level need not produce identical scores under different class prototypes. To study the effect, consider the abstract decomposition

$$Z = U_\mathcal{S} + V_\mathcal{S}, \tag{7}$$

where $U_\mathcal{S}$ carries the evidence shared among the classes in $\mathcal{S}$ and $V_\mathcal{S}$ is the remaining candidate-discriminative residual. How these are estimated is deferred to Section 3. For the ground-truth class y and a competitor $c \in \mathcal{S}$, the pairwise margin splits as

$$\underbrace{(w_y - w_c)^\top Z}_{m_Z(y,c)} = \underbrace{(w_y - w_c)^\top V_\mathcal{S}}_{m_V(y,c)} + \underbrace{(w_y - w_c)^\top U_\mathcal{S}}_{b_U(y,c)}, \tag{8}$$

with m_V the margin from the residual and b_U the margin the shared component adds. Writing each candidate's response to $U_{\mathcal{S}}$ as the candidate average $\overline{s}^U_{\mathcal{S}}$ plus a deviation $e_{c,\mathcal{S}}$, the average cancels from every pairwise comparison and only the deviations survive:

$$b_U(y, c) = e_{y,\mathcal{S}} - e_{c,\mathcal{S}}. \tag{9}$$

Global training does not force $e_{c,\mathcal{S}} = 0$ for every class and every input-dependent $\mathcal{S}$, so concept-level sharedness does not guarantee score-level cancellation.

Proposition 2 (Dominance amplifies imperfect local cancellation)**.** *For any $y, c \in \mathcal{S}$,*

$$|b_U(y, c)| \leq \|w_y - w_c\|_2 \, \|U_{\mathcal{S}}\|_2. \tag{10}$$

If the residual favors y but the shared component opposes it strongly enough,

$$m_V(y, c) > 0 \quad \textit{and} \quad b_U(y, c) < -m_V(y, c), \tag{11}$$

then $m_Z(y, c) < 0$: the full representation reverses a decision the residual gets right, and removing $U_{\mathcal{S}}$ restores it.

Proof in Appendix D. Proposition 2 is a conditional failure mode. Even when shared evidence is nearly common, a small class-specific deviation, scaled by a large $\|U_{\mathcal{S}}\|$, can outweigh a weak fine-grained margin. Figure 2 makes the magnitude condition plausible: the atoms most uniformly prevalent across probable classes are also the most strongly expressed. It does not show that $U_{\mathcal{S}}$ dominates every representation or harms every prediction. Whether interference occurs for a given example depends on both an adverse deviation and a large enough magnitude relative to the residual margin; both are measured directly in Section 4.

3 Candidate-Restricted Reranking

To examine the global–local mismatch in Section 2, we study two forms of candidate-restricted reranking. The first operates on the full representation Z, testing candidate-local optimization alone. The second uses an SAE-based decomposition to explicitly remove candidate-shared concepts before an otherwise identical reranker. Both keep the visual encoder and first-stage classifier frozen and only reorder the retrieved top-(K) candidates. Figure 3 provides an overview.

3.1 Reranking on the Full Representation

Proposition 1 shows that global training rewards evidence that separates the candidate group from non-candidates even when it gives no preference within the group. The direct fix is to restrict the second-stage objective to the probable classes. After retrieving $\mathcal{S}(Z)$ with the frozen global classifier, we attach a linear reranker $r_\phi(Z) = W_\phi Z + b_\phi \in \mathbb{R}^C$ and train it with cross-entropy normalized over the candidates only:

$$\ell_{\text{rerank}}(Z, y) = -r_{\phi,y}(Z) + \log \sum_{c \in \mathcal{S}(Z,y)} \exp r_{\phi,c}(Z). \tag{12}$$

Here $\mathcal{S}(Z, y)$ is a label-aware training set: it equals $\mathcal{S}(Z)$ when $y \in \mathcal{S}(Z)$, and otherwise replaces the lowest-ranked candidate with y so the target is always present. At inference, the natural set $\mathcal{S}(Z)$ is used with no label access. Only ϕ is trained.

3.2 Reranking via Explicit Candidate-Conditioned Intervention

The adaptation in Section 3.1 is unconstrained, however. Since the complete Z is still available and shared concepts are still present, the reranker may still exploit correlations involving them, and any reduced reliance cannot be inspected. To enable a controlled mechanistic study, we therefore introduce an explicit intervention via SAE decomposition (see Appendix B for more discussion).

Residual sufficiency. Recall $Z = U_{\mathcal{S}} + V_{\mathcal{S}}$ from Equation 7 and the retrieval event $\mathcal{E}_K = \{Y \in \mathcal{S}(Z)\}$. Since prevalence is only an estimator of sharedness, we state the ideal condition:

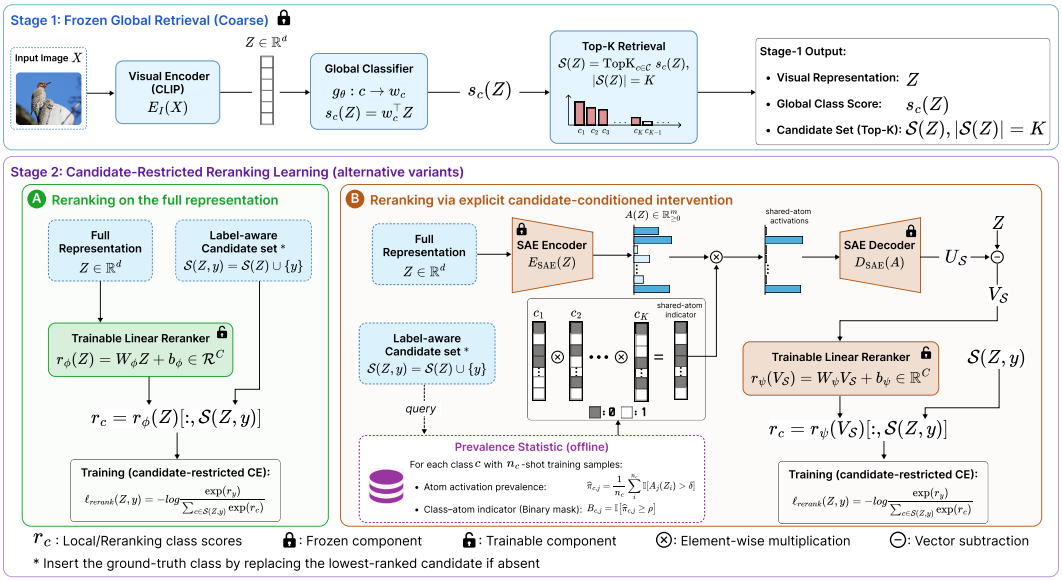


Figure 3: **Overview of candidate-restricted reranking learning.** The frozen encoder and global classifier retrieve the top-K candidate set $\mathcal{S}(Z)$. **(A)** Reranking on the full representation adapts implicitly to the local objective. **(B)** Reranking via explicit candidate-conditioned intervention identifies, decodes, and removes candidate-shared concepts, then reranks on the SAE residual.

Assumption 1 (Candidate-conditioned invariance)**.** Conditioned on successful retrieval, the shared component carries no class information beyond what the residual and candidate set already provide:

$$Y \perp U_{\mathcal{S}} \mid V_{\mathcal{S}}, \mathcal{S}, \mathcal{E}_K. \tag{13}$$

Theorem 1 (Candidate-level residual sufficiency)**.** *Under Assumption 1,*

$$p(Y \mid Z, \mathcal{S}, \mathcal{E}_K) = p(Y \mid V_{\mathcal{S}}, \mathcal{S}, \mathcal{E}_K). \tag{14}$$

Hence $V_{\mathcal{S}}$ is sufficient for predicting the class within $\mathcal{S}$, and the Bayes risks from Z and $V_{\mathcal{S}}$ coincide on $\mathcal{E}_K$.

Proof in Appendix E, together with the equivalent mutual-information statement (Corollary 2). The theorem says removal is safe under the assumption; it does not predict a gain, since a Bayes-optimal classifier could already ignore $U_{\mathcal{S}}$.

Candidate-conditioned shared-concept estimation. Using the class prevalence $\widehat{\pi}_{c,j}$ of Equation 3, computed on the few-shot training set, we mark atom j as characteristic of class c when its prevalence clears a threshold ρ, and as shared by the current candidate set when it is characteristic of every candidate:

$$B_{c,j} = \mathbb{I}[\widehat{\pi}_{c,j} > \rho], \qquad M_{\mathcal{S},j} = \prod_{c \in \mathcal{S}} B_{c,j}. \tag{15}$$

Sharedness is decision-dependent: the same atom is removed for one candidate set and kept for another. Common prevalence is a proxy for Assumption 1, not a guarantee. We discuss such limitation in Appendix H and the motivation for SAE usage in Appendix B.

Explicit residualization and reranking. For a query Z, we decode only the masked activations and subtract them:

$$U_{\mathcal{S}} = D_{\mathrm{SAE}}(M_{\mathcal{S}} \odot A(Z)), \qquad V_{\mathcal{S}} = Z - U_{\mathcal{S}}. \tag{16}$$

Defining $V_{\mathcal{S}}$ by subtraction makes Equation 7 exact regardless of SAE reconstruction error. An identical linear reranker $r_\psi(V_{\mathcal{S}})$ is then trained with the same candidate-restricted loss Equation 12, with $V_{\mathcal{S}}$ in place of Z and the mask built from $\mathcal{S}(Z, y)$ during training. The two formulations share the candidate set, architecture, and objective, and differ only in their input. Once the encoder, SAE, global classifier, prevalence statistics, and thresholds are fixed, the map from $(Z, \mathcal{S})$ to $V_{\mathcal{S}}$ is deterministic and reproducible.

Table 1: Main results on 11 image-classification datasets using CLIP–ResNet-50 in the 16-shot, top-15 settings. Entries are top-1 accuracy Acc_h (%). **Bold** and underlined values denote the best and second-best results. Please refer to Table 2 in the appendix for 8- and 4-shot settings.

Global classifier	Reranker input	Dataset											Avg. (Acc_h)
		ImageNet	CUB-200	DTD	Flowers102	StanfordCars	OxfordPets	CIFAR100	FGVCAircraft	Food101	Caltech101	UCF101	
CLIP (zero-shot)		55.49	46.68	42.39	66.43	54.17	85.66	38.98	17.07	80.59	83.16	57.81	57.13
Linear Probe	None	56.14	65.97	66.70	92.69	69.99	76.80	50.96	35.03	74.32	91.19	74.42	68.56
	Z	**58.36**	**67.58**	**67.60**	92.73	**71.88**	**78.00**	52.33	**35.94**	76.21	**92.11**	**75.96**	**69.88**
	V_S	58.20	67.35	67.51	**92.79**	71.35	77.75	**52.58**	35.92	**76.25**	92.08	**75.96**	69.79
CoOp	None	60.30	66.33	64.64	91.51	72.00	89.63	53.89	32.99	82.66	91.99	76.48	71.13
	Z	**64.23**	**70.07**	**66.50**	92.74	75.24	**90.16**	55.49	35.66	**83.23**	**93.33**	**78.77**	**73.22**
	V_S	64.16	70.04	66.48	**92.81**	**75.25**	90.05	**55.55**	**35.77**	83.21	93.30	78.71	73.21
CoCoOp	None	60.93	57.13	59.41	80.94	64.58	89.57	50.11	20.12	82.58	91.41	72.67	66.31
	Z	**64.77**	**69.47**	**65.03**	89.55	**75.05**	90.16	54.91	31.07	83.22	**93.11**	**78.42**	**72.25**
	V_S	64.70	**69.47**	64.76	**89.61**	74.63	**90.17**	**55.26**	**31.26**	**83.24**	93.10	78.28	72.23

3.3 Global–Local Evidence Fusion

The rerankers specialize in candidate-local discrimination, while the frozen global classifier holds the broader evidence that identified the plausible classes. Replacing the global prediction outright would discard that signal. We therefore keep the global scores and let the reranker adjust the ordering. Writing r_c for the local logit of either variant, for each $c \in \mathcal{S}(Z)$ we form

$$h_c^{(\lambda)}(Z) = s_c(Z) + \lambda r_c, \qquad \lambda \geq 0, \tag{17}$$

and predict $\arg\max_{c \in \mathcal{S}(Z)} h_c^{(\lambda)}(Z)$. $\lambda = 0$ recovers the global ordering; $\lambda = 1$ is an unweighted sum, equivalent to a reranker temperature $\tau_r = 1/\lambda$. This logit fusion strategy has an exact probabilistic interpretation rather than a heuristic choice. The full discussion are in Appendix E.5.

4 Experiments

4.1 Experiment Settings

Datasets and global classifier baselines. Our study is conducted on eleven image-classification benchmarks, including ImageNet (Deng et al., 2009), CUB-200 (Wah et al., 2011), DTD (Cimpoi et al., 2014), Flowers102 (Nilsback & Zisserman, 2008), StanfordCars (Krause et al., 2013), OxfordPets (Parkhi et al., 2012), CIFAR100 (Krizhevsky et al., 2009), FGVCAircraft (Maji et al., 2013), Food101 (Bossard et al., 2014), Caltech101 (Fei-Fei et al., 2004), and UCF101 (Soomro et al., 2012), on few-shot settings. We instantiate the frozen first-stage global classifier using three common adaptation strategies: a linear classifier and two prompt learning methods, including CoOp and CoCoOp (Zhou et al., 2022b;a). Detailed training configurations for the three global classifiers and dataset characteristics are provided in Appendix F.

Training settings. Our main experiments use CLIP with a ResNet-50 visual encoder (Radford et al., 2021). To examine whether the findings transfer across encoder families, we additionally report results with CLIP ViT-B/32 in Appendix G. Our training procedures for the reranking heads are detailed in Appendix F. The concept decomposition uses the TopK SAE variant described in Section 2 and Appendix B.

Evaluation metrics. We use top-1 classification accuracy as the evaluation metric. For the overall performance of the whole system, we report the accuracy using hybrid logits h_c (Equation 17), denoted Acc_h, and the reranking accuracy using only the reranking logit r_c, denoted Acc_r. All reported results are averaged over three different random seed runs.

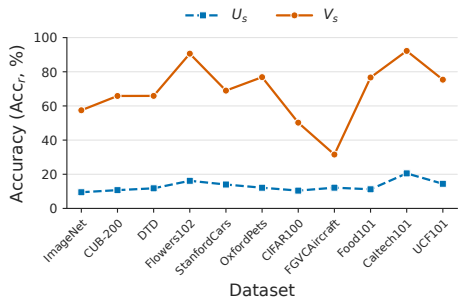


Figure 4: Standalone reranking accuracy on (U_S) versus (V_S). Stage-1 is Linear Probe.

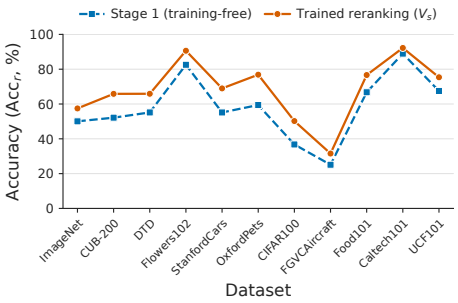


Figure 5: Training-free versus trained reranking on the residual representation.

4.2 FINDINGS

Candidate-restricted reranking improves every global classifier. Table 1 shows that candidate-restricted reranking improves the macro-average accuracy of all three global classifiers at 16 shots, whether the reranker receives the full representation Z or the SAE residual V_S. Reranking on Z improves Linear Probe, CoOp, and CoCoOp by 1.32, 2.09, and 5.94 percentage points, respectively. This consistency across classifier types suggests that an additional candidate-local discrimination step can refine decisions made by the global classifier. The improvement from full-Z reranking shows that local refinement is effective even when the reranker retains access to every component of the original representation. Explicitly suppressing the SAE-selected shared component yields comparable results: its macro-average differs from full-Z reranking by at most 0.09 percentage points. Together with the prevalence analysis in Section 2, this comparison gives the SAE a concrete diagnostic role: it identifies which atoms are treated as candidate-shared, while subtraction tests the predictive effect of withholding their decoded contributions from the local head. The comparable fused accuracy indicates that this explicit intervention can preserve candidate-local refinement. Because the frozen global scores still use Z, the result concerns the evidence available to the reranker rather than the necessity of these components to the entire prediction pipeline.

The removed component is large but carries little candidate-local information. Candidate-shared should not be conflated with either negligible or semantically empty. We therefore examine both the magnitude of U_S and its ability to support candidate-local classification. As shown in Figure 1a, the ratio $\|U_S\|_2/\|Z\|_2$ ranges from 0.61 to 0.78 across the eleven datasets, with an average of 0.68. The decoded shared component therefore constitutes a noticeable part of the original representation rather than a small residual artifact. Despite its magnitude, U_S provides limited candidate-local evidence. In Figure 4, reranking on V_S consistently outperforms reranking on U_S, achieving (68.33%) versus (13.00%) average standalone accuracy. The prevalence-based intervention thus removes a high-magnitude component that retains some predictive signal but is far less useful for distinguishing among the current candidates. Conversely, the residual V_S concentrates the evidence required for candidate-local discrimination. Together with the comparable hybrid accuracy of full-Z and residual reranking, this result shows that withholding U_S from the local head does not materially compromise prediction. Because the global classifier still operates on Z, U_S may remain useful for global retrieval; its limited value is specific to the candidate-local decision.

Candidate-restricted training suppresses the influence of shared evidence. Proposition 2 shows that candidate-shared evidence can alter local decisions when its class-specific score contributions do not cancel. For each test example with true class y, we select its strongest competing candidate $c^\star$ under the stage-1 classifier and measure $b_U^{(h)}(y, c^\star) = (w_y^{(h)} - w_{c^\star}^{(h)})^\top U_S$, where h denotes the global classifier or the full-Z reranker. Holding U_S and the class pair fixed isolates the classifier-head effect. We average $|b_U^{(h)}|$ within each ground-truth class and then across classes for Figure 1c, with classwise results in Figure 10 in the appendix. The global classifier exhibits substantial deviations (0.82–3.04), whereas full-Z reranking reduces them to 0.10–0.56, lowering the average from 1.28 to 0.33—approximately 74%. Together with the substantial magnitude of U_S in Figure 1a, these findings support the imperfect-cancellation mechanism in Proposition 2: promi-

nent shared evidence produces non-negligible candidate-specific score deviations. Candidate-local training suppresses this influence implicitly, while residualization makes the removed component's direct contribution zero by construction.

Residual evidence is informative, and local training improves its use. To separate the information in V_S from the effect of candidate-local optimization, we compare two scoring rules while holding the candidate set and residual fixed. The training-free variant applies the frozen stage-1 head directly to V_S, whereas the trained variant uses the candidate-restricted reranker. This comparison tests whether the globally trained decision rule can directly exploit the residual evidence or requires local adaptation. As shown in Figure 5, the frozen head achieves 58.14% average accuracy on V_S, demonstrating that the residual retains discriminative directions accessible to the global classifier. The trained reranker improves every dataset, reaching 68.32% on average—a gain of 10.18 percentage points. Thus, residualization exposes meaningful candidate-discriminative evidence, while candidate-restricted training learns to use it more effectively within the local decision context.

Please refer to Appendix G for more ablation studies, qualitative examples, and discussion.

5 Related Works

Task-relevant information in deep representations. Representation learning seeks transformations that expose factors useful for downstream prediction and transfer (Bengio et al., 2013). Information-theoretic accounts formalize this goal by preserving target-relevant information while compressing the input and relating minimal sufficiency to invariance against task-defined nuisances (Tishby et al., 2000; Achille & Soatto, 2018). Thus, relevance and redundancy are task-relative, consistent with evidence that feature generality depends on network depth, task similarity, and pretraining (Yosinski et al., 2014; Kornblith et al., 2019). CLIP-like models extend this transfer paradigm through broad image–text pretraining (Radford et al., 2021; Zhai et al., 2023), but their frozen representations need not be minimal for a particular downstream decision. We therefore study how the role of a representation component changes when prediction is narrowed from the complete label space $\mathcal{C}$ to an input-dependent candidate set $\mathcal{S}(Z)$, where globally useful evidence may become non-discriminative among probable alternatives.

Vision–Language adaptation and candidate-aware prediction. CLIP supports zero-shot transfer and efficient downstream adaptation through its frozen visual features (Radford et al., 2021). Among lightweight adaptation methods, CoOp learns continuous context tokens from few labeled examples, while CoCoOp generates input-conditional prompts to improve generalization beyond training classes (Zhou et al., 2022b;a). To distinguish ambiguous classes, FuDD (Esfandiarpoor & Bach, 2024) generates differential descriptions using large language models, while CODER (Yi et al., 2024) represents images through image–text neighbor distances and reranks probable classes using pairwise descriptions. PCNN (Chen et al., 2023) likewise refines predictions within retrieved candidates, with accuracy bounded by first-stage recall. LDC (Li et al., 2025) and CAPT (Shao et al., 2026) address inter-class confusion through logit refinement and confusion-aware prompt tuning, respectively. Complementing these approaches, we provide a mechanistic study on how the relevance of frozen representation components changes after conditioning on probable classes, combining an analysis of global–local objective mismatch with matched reranking experiments to test whether suppressing candidate-shared components preserves local predictive performance.

6 Conclusion

We studied how the relevance of information in a frozen CLIP representation changes when the prediction scope narrows from the full label space to the candidate classes a global classifier retrieves. Using a sparse autoencoder, we found that concepts prevalent across candidate classes are also strongly expressed, and we gave an objective-level account of why a globally trained classifier retains them: the global cross-entropy rewards evidence that separates the candidate group from the rest of the label space, even when that evidence cancels within the group. Aligning the objective with the local decision improves accuracy. Explicitly suppressing the candidate-shared component matches this, identifying what the restricted objective learns to discount and showing that roughly two-thirds of the representation by norm is dispensable once the candidate set is fixed.

AI USE STATEMENT

We used generative AI tools (large language models) to aid and polish the writing of this paper, including rephrasing sentences, tightening prose, and checking consistency of notation across sections. We also used them to assist with plotting code for the figures. We did not use generative AI tools to generate research ideas, design experiments, produce or analyze experimental results, or write proofs; all theoretical statements and their proofs were derived and verified by the authors. All AI-assisted text and code were reviewed, edited, and tested by the authors, and we take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies the internal representations of a publicly available vision foundation model on standard, publicly available image-classification benchmarks. It involves no human subjects, no new data collection, and no personally identifiable information. We do not foresee direct negative societal impact; the analysis is diagnostic and does not introduce new deployment capabilities.

REPRODUCIBILITY STATEMENT

All experiments use public datasets and public pretrained CLIP weights. The method is specified in Section 3; hyperparameters, training schedules, and the SAE architecture are given in Appendix B and Appendix F. Every theoretical statement is proved in Appendix D and Appendix E, with all assumptions stated explicitly. Reported accuracies are means over three fixed seeds (0, 42, 123). Code and the frozen SAE checkpoint will be released upon publication.

REFERENCES

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063, 2023.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). *Advances in Neural Information Processing Systems*, 37:84298–84328, 2024.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Valerie Chen, Mohammad Reza Taesiri, Anh Totti Nguyen, et al. Pcnn: probable-class nearest-neighbor explanations improve fine-grained image classification accuracy for ais and humans. *arXiv preprint arXiv:2308.13651*, 2023.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Reza Esfandiarpoor and Stephen Bach. Follow-up differential descriptions: Language models resolve ambiguities for image classification. In *International Conference on Learning Representations*, volume 2024, pp. 5849–5873, 2024.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

Thomas Fel, Ekdeep Singh Lubana, Jacob S. Prince, Matthew Kowal, Victor Boutin, Isabel Papadimitriou, Binxu Wang, Martin Wattenberg, Demba E. Ba, and Talia Konkle. Archetypal SAE: Adaptive and stable dictionary learning for concept extraction in large vision models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=9v1eW8HgMU.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pp. 26721–26754, 2025.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Sonia Joseph, Praneet Suresh, Ethan Goldfarb, Lorenz Hufe, Yossi Gandelsman, Robert Graham, Danilo Bzdok, Wojciech Samek, and Blake Aaron Richards. Steering clip's vision transformer with sparse autoencoders. *arXiv preprint arXiv:2504.08729*, 2025.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2661–2671, 2019.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *International Conference on Learning Representations*, volume 2025, pp. 53617–53642, 2025.

Shuo Li, Fang Liu, Zehua Hao, Xinyi Wang, Lingling Li, Xu Liu, Puhua Chen, and Wenping Ma. Logits deconfusion with clip for few-shot learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25411–25421. IEEE, 2025.

Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. In *International Conference on Learning Representations*, volume 2025, pp. 24444–24469, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Maoyuan Shao, Yutong Gao, Xinyang Huang, Lijuan Sun, Guoshun Nan, and Chuang Zhu. Capt: Confusion-aware prompt tuning for reducing vision-language misalignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3154–3164, 2026.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Chao Yi, Lu Ren, De-Chuan Zhan, and Han-Jia Ye. Leveraging cross-modal neighbor representation for improved clip classification. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27392–27401. IEEE, 2024.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

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

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 16795–16804. IEEE, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022b.

Appendix

Appendix Contents.

- **Extended Related Works** (Section A). Reviews methods for interpreting and intervening on vision foundation representations, including sparse concept decomposition, SAE-based analysis and steering, and concept erasure, while clarifying how SAE atoms are treated in this work.
- **SAE Architecture and Training** (Section B). Describes the TopK sparse autoencoder architecture, training objective, hyperparameters, and computational setup, and explains how concept decomposition enables mechanistic diagnosis and candidate-conditioned interventions.
- **Prevalence–Magnitude Sharedness Diagnostic** (Section C). Defines the probable class sets, prevalence- and magnitude-based sharedness scores, valid-atom selection procedure, and class-balanced correlation analysis used to measure candidate-shared structure.
- **Proofs for Section 2** (Section D). Provides proofs and extended derivations for the global–local objective mismatch, including the global incentive to preserve candidate-shared evidence, the decomposition of its score contribution, and the conditions under which it can reverse a candidate-local margin.
- **Proofs and Extended Statements for Section 3** (Section E). Establishes the sufficiency of the residual representation under the stated conditional-invariance assumption, gives its information-theoretic form, specifies the label-aware reranking objective, and derives the probabilistic and variational interpretations of global–local logit fusion.
- **Details of Experiment Settings** (Section F). Describes the eleven evaluation datasets, the Linear Probe, CoOp, and CoCoOp global classifiers, and the training configurations used for the candidate-restricted rerankers.
- **More Experimental Results** (Section G). Reports complete results for CLIP–ResNet-50 and CLIP–ViT-B/32 across shot settings, studies candidate-set size and prevalence threshold, and presents classwise analyses and qualitative examples of successful and unsuccessful reranking.
- **Limitations and Broader Impact** (Section H). Discusses the dependence on successful candidate retrieval and the approximate shared-atom selection rule, then outlines how candidate conditioning may support the analysis and design of coarse-to-fine prediction systems across tasks and modalities.

A Extended Related Works

Interpreting and intervening on vision foundation representations. Neural networks can encode multiple semantically distinct factors within the same activation coordinates, a phenomenon known as *polysemanticity* (Elhage et al., 2022; Huben et al., 2024). This entanglement makes individual dimensions of dense representations difficult to interpret as distinct semantic factors. Sparse autoencoders (SAEs) (Huben et al., 2024) address this difficulty by reconstructing dense activations from a sparse, typically overcomplete dictionary of latent directions, providing a more inspectable basis for component-level analysis. Recent work has extended sparse concept decomposition to vision foundation models: SpLiCE (Bhalla et al., 2024) represents CLIP embeddings as sparse combinations of text-defined concepts; PatchSAE (Lim et al., 2025) extracts spatially localized concepts, including shape, color, and object semantics, and analyzes their use during prompt adaptation, finding that much of the performance gain can be explained by concepts already present in the pretrained model; and hierarchical and archetypal SAEs recover concepts at multiple granularities or improve dictionary stability in large vision models (Zaigrajew et al., 2025; Fel et al., 2025). SAE features have also supported targeted steering and suppression of CLIP behavior (Joseph et al., 2025). Related concept-erasure work studies interventions that remove specified information from representations. In particular, LEACE prevents linear classifiers from detecting a target concept while minimizing changes to the representation (Belrose et al., 2023). Because SAE atoms are not guaranteed to form unique or canonical units of analysis (Leask et al., 2025), we treat them as operational representation components rather than ground-truth concepts. In our study, the SAE serves as an analysis instrument for changes in prediction scope: we condition atom sharedness on the probable class set and test whether candidate-local predictive performance can be preserved after suppressing selected atoms' decoded contributions in an architecture-matched intervention.

B SAE Architecture and Training

We use a TopK sparse autoencoder (Gao et al., 2025). Sparsity is enforced by retaining only the k_{SAE} largest encoder activations,

$$\|E_{\mathrm{SAE}}(Z)\|_0 \le k_{\mathrm{SAE}}. \tag{18}$$

The encoder and decoder dictionaries are trained without class labels by minimizing reconstruction error under this constraint:

$$\min_{E_{\mathrm{SAE}},D} \mathbb{E}_Z\left[\|Z - DE_{\mathrm{SAE}}(Z)\|_2^2\right] \quad \text{subject to} \quad \|E_{\mathrm{SAE}}(Z)\|_0 \le k_{\mathrm{SAE}}. \tag{19}$$

SAE training. We train a TopK sparse autoencoder on ImageNet representations extracted from the frozen CLIP–ResNet-50 or CLIP–ViT-B/32 visual encoder. The SAE uses an expansion factor of 8, retaining the 32 largest latent activations per input. Training runs for 100 epochs with a batch size of 2048, a learning rate of 5×10^{-4}, 500 warm-up steps, and no weight decay. We use AdamW (Loshchilov & Hutter, 2019) as the optimizer. The trained SAE is then frozen and reused across downstream datasets. All SAE training is conducted on a single NVIDIA B200 180GB.

Why concept decomposition? Candidate-restricted training changes the optimization scope, but its adaptation is implicit: it does not reveal which components of Z are ignored or whether shared evidence is actually suppressed. Concept decomposition makes the hypothesis testable. **Mechanistic diagnosis.** Inspectable atoms let us measure which concepts recur across candidates, how strongly they are expressed, and how their score contributions relate to local interference. **Candidate-conditioned decomposition.** It splits Z into $U_{\mathcal{S}}$ and $V_{\mathcal{S}}$ according to the current candidate set, reflecting that a concept's value depends on the decision context. **Direct intervention.** With all components frozen, subtracting $U_{\mathcal{S}}$ is deterministic and denies the reranker access to the selected evidence. **Transparency.** The removed atoms and the criterion that selected them can be inspected, so each step of the mechanism can be evaluated on its own. The SAE turns implicit adaptation into an explicit, auditable intervention.

C Prevalence–Magnitude Sharedness Diagnostic

This appendix gives the scores behind Figure 2. Prevalence $\widehat{\pi}_{c,j}$ and magnitude $\widehat{\mu}_{c,j}$ are defined in Equation 3.

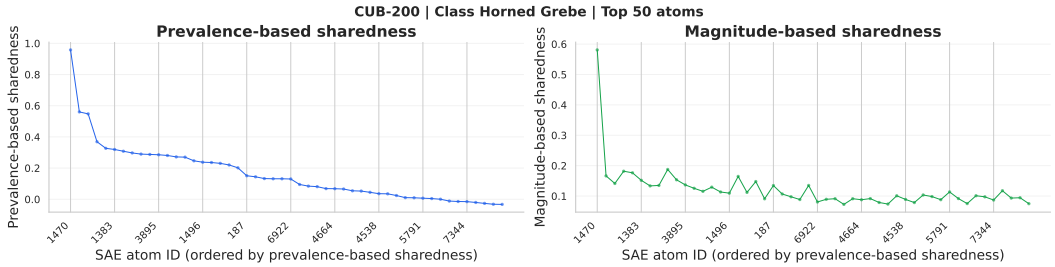


Figure 6: Atom-level profiles for CUB-200 class 50 (Horned Grebe). The same 50 atoms appear in both plots, ordered by decreasing prevalence-based sharedness.

Class-level probable set. For each target class c we build a probable set $\mathcal{S}_c$ containing c and the $K-1$ classes that appear most often in the global classifier's top-K predictions on examples of class c. This gives a stable summary of the classes c is commonly confused with, independent of any single image.

Sharedness scores. For atom j within $\mathcal{S}_c$ we define

$$Q^P_{\mathcal{S}_c,j} = \mathrm{Mean}_{c'\in\mathcal{S}_c}(\widehat{\pi}_{c',j}) - \mathrm{Std}_{c'\in\mathcal{S}_c}(\widehat{\pi}_{c',j})\,, \tag{20}$$

$$Q^M_{\mathcal{S}_c,j} = \mathrm{Mean}_{c'\in\mathcal{S}_c}(\widehat{\mu}_{c',j}) - \mathrm{Std}_{c'\in\mathcal{S}_c}(\widehat{\mu}_{c',j})\,. \tag{21}$$

The mean rewards atoms expressed broadly across the probable classes; the standard deviation penalizes class-specific variation. Q^P measures sharedness through occurrence, Q^M through strength.

Valid atoms. Atoms that barely fire in the neighborhood have poorly estimated class statistics, so we keep only the N atoms with the highest mean prevalence across $\mathcal{S}_c$:

$$f_{\mathcal{S}_c,j} = \frac{1}{|\mathcal{S}_c|}\sum_{c'\in\mathcal{S}_c}\widehat{\pi}_{c',j}, \qquad \mathcal{J}_c = \mathrm{TopN}_j\ f_{\mathcal{S}_c,j}. \tag{22}$$

Correlation and aggregation. For each target class, we compute the Pearson correlation over valid atoms,

$$r_c = \mathrm{Corr}_{j\in\mathcal{J}_c}\left(Q^P_{\mathcal{S}_c,j},\, Q^M_{\mathcal{S}_c,j}\right). \tag{23}$$

Because correlations are bounded and their sampling distribution need not be symmetric, we summarize a dataset by averaging class-wise correlations on the Fisher z scale,

$$\overline{r}_{\mathrm{F}} = \tanh\left(\frac{1}{C_{\mathrm{valid}}}\sum_{c\in\mathcal{C}_{\mathrm{valid}}}\operatorname{arctanh}(r_c)\right). \tag{24}$$

Equal weighting makes the class, rather than the class–atom pair, the unit of analysis.

Figure 6 illustrates a representative example of how atom prevalence and its magnitude sharedness correlate in the CUB-200 dataset.

D Proofs for Section 2

D.1 Proof of Proposition 1

Let

$$A^{(0)}_{\mathcal{S}} = \sum_{c\in\mathcal{S}}\exp s^{(0)}_c(Z), \qquad B^{(0)}_{\mathcal{S}} = \sum_{r\notin\mathcal{S}}\exp s^{(0)}_r(Z). \tag{25}$$

Under the shift in Equation 6, the same score is added to every candidate logit, so it cancels from the restricted softmax:

$$\begin{aligned}\ell^{(\alpha)}_{\mathcal{S}}(Z,y) &= -s^{(0)}_y(Z) - \alpha + \log\left(e^{\alpha}A^{(0)}_{\mathcal{S}}\right)\\ &= -s^{(0)}_y(Z) + \log A^{(0)}_{\mathcal{S}}.\end{aligned} \tag{26}$$

Hence, the common score does not change any relative probability within $\mathcal{S}$:

$$\ell_{\mathcal{S}}^{(\alpha)}(Z, y) = \ell_{\mathcal{S}}^{(0)}(Z, y). \tag{27}$$

The global loss becomes

$$\ell_{\mathcal{C}}^{(\alpha)}(Z, y) = \ell_{\mathcal{S}}^{(0)}(Z, y) + \log\left(1 + e^{-\alpha}\frac{B_{\mathcal{S}}^{(0)}}{A_{\mathcal{S}}^{(0)}}\right). \tag{28}$$

If $\mathcal{C} \setminus \mathcal{S}$ is nonempty then $B_{\mathcal{S}}^{(0)} > 0$ and the second term is strictly decreasing in α. □

D.2 Probabilistic Form of the Training Incentive

Corollary 1 (Probabilistic form of the training incentive)**.** *Let $p_\theta^{(\alpha)}$ be the global predictive distribution induced by the shifted scores in Equation 6. Then*

$$\frac{\partial \ell_{\mathcal{C}}^{(\alpha)}(Z, y)}{\partial \alpha} = -p_\theta^{(\alpha)}(Y \notin \mathcal{S} \mid Z) = -\sum_{r \notin \mathcal{S}} p_\theta^{(\alpha)}(Y = r \mid Z). \tag{29}$$

The incentive to raise the common candidate score equals the total probability mass the classifier assigns outside the candidate set.

Proof. Under the shifted scores,

$$p_\theta^{(\alpha)}(Y \notin \mathcal{S} \mid Z) = \frac{B_{\mathcal{S}}^{(0)}}{e^\alpha A_{\mathcal{S}}^{(0)} + B_{\mathcal{S}}^{(0)}}. \tag{30}$$

Differentiating Equation 28 with respect to α gives the negative of this quantity. □

The incentive is strongest when the classifier places substantial mass outside $\mathcal{S}$, where raising a shared score transfers mass efficiently from non-candidates to the candidate group. It weakens as that mass approaches zero, but global training never introduces an opposing incentive to remove shared evidence on the grounds that it leaves within-$\mathcal{S}$ probabilities unchanged.

D.3 Score Decomposition Behind Equation 9

Write the shared component's score for candidate c as its candidate average plus a deviation,

$$w_c^\top U_{\mathcal{S}} = \overline{s}_{\mathcal{S}}^{U} + e_{c,\mathcal{S}}, \qquad \overline{s}_{\mathcal{S}}^{U} = \frac{1}{|\mathcal{S}|}\sum_{r \in \mathcal{S}} w_r^\top U_{\mathcal{S}}. \tag{31}$$

Substituting into $b_U(y, c) = (w_y - w_c)^\top U_{\mathcal{S}}$, the average cancels and $b_U(y, c) = e_{y,\mathcal{S}} - e_{c,\mathcal{S}}$, which is Equation 9.

D.4 Proof of Proposition 2

The bound Equation 10 is Cauchy–Schwarz applied to $b_U(y, c) = (w_y - w_c)^\top U_{\mathcal{S}}$. For the reversal, combine Equation 8 with the condition Equation 11:

$$m_Z(y, c) = m_V(y, c) + b_U(y, c) < 0, \tag{32}$$

while removing $U_{\mathcal{S}}$ leaves $m_V(y, c) > 0$. □

E Proofs and Extended Statements for Section 3

E.1 Density Form of Assumption 1

The conditional independence in Equation 13 is equivalent to: for every $c \in \mathcal{S}$,

$$p(U_{\mathcal{S}} \mid V_{\mathcal{S}}, Y = c, \mathcal{S}, \mathcal{E}_K) = p(U_{\mathcal{S}} \mid V_{\mathcal{S}}, \mathcal{S}, \mathcal{E}_K). \tag{33}$$

E.2 Proof of Theorem 1

For a fixed realized candidate set $\mathcal{S}$, both $U_{\mathcal{S}}$ and $V_{\mathcal{S}}$ are deterministic functions of Z, and Z is recovered from $(U_{\mathcal{S}}, V_{\mathcal{S}})$ through Equation 7. Conditioning on Z is therefore equivalent to conditioning on $(U_{\mathcal{S}}, V_{\mathcal{S}})$: for every $c \in \mathcal{S}$,

$$p(Y = c \mid Z, \mathcal{S}, \mathcal{E}_K) = p(Y = c \mid U_{\mathcal{S}}, V_{\mathcal{S}}, \mathcal{S}, \mathcal{E}_K). \tag{34}$$

By conditional Bayes' rule,

$$\begin{aligned} &p(Y = c \mid U_{\mathcal{S}}, V_{\mathcal{S}}, \mathcal{S}, \mathcal{E}_K) \\ &\quad = \frac{p(U_{\mathcal{S}} \mid V_{\mathcal{S}}, Y = c, \mathcal{S}, \mathcal{E}_K)\, p(Y = c \mid V_{\mathcal{S}}, \mathcal{S}, \mathcal{E}_K)}{p(U_{\mathcal{S}} \mid V_{\mathcal{S}}, \mathcal{S}, \mathcal{E}_K)}. \end{aligned} \tag{35}$$

By Equation 33 the numerator's likelihood equals the denominator, so

$$p(Y = c \mid U_{\mathcal{S}}, V_{\mathcal{S}}, \mathcal{S}, \mathcal{E}_K) = p(Y = c \mid V_{\mathcal{S}}, \mathcal{S}, \mathcal{E}_K). \tag{36}$$

Combining Equation 34 and Equation 36 gives Equation 14. Since the full candidate posterior is unchanged, so are its maximizer and the conditional Bayes risk. □

E.3 Information-Theoretic Form

Corollary 2 (Information-theoretic form of residual sufficiency)**.** *Under Assumption 1,*

$$I(Y; Z \mid \mathcal{S}, \mathcal{E}_K) = I(Y; V_{\mathcal{S}} \mid \mathcal{S}, \mathcal{E}_K), \tag{37}$$

and equivalently

$$H(Y \mid Z, \mathcal{S}, \mathcal{E}_K) = H(Y \mid V_{\mathcal{S}}, \mathcal{S}, \mathcal{E}_K). \tag{38}$$

Proof. For fixed $\mathcal{S}$, Z and $(U_{\mathcal{S}}, V_{\mathcal{S}})$ determine each other. The chain rule for conditional mutual information gives

$$\begin{aligned} I(Y; Z \mid \mathcal{S}, \mathcal{E}_K) &= I(Y; U_{\mathcal{S}}, V_{\mathcal{S}} \mid \mathcal{S}, \mathcal{E}_K) \\ &= I(Y; V_{\mathcal{S}} \mid \mathcal{S}, \mathcal{E}_K) + I(Y; U_{\mathcal{S}} \mid V_{\mathcal{S}}, \mathcal{S}, \mathcal{E}_K). \end{aligned} \tag{39}$$

The last term is zero under Assumption 1, which proves Equation 37. Writing mutual information as a reduction in conditional entropy gives Equation 38. □

E.4 Residual Reranker and Label-Aware Candidate Set

For completeness, the residual reranker and its objective are

$$r_\psi(V_{\mathcal{S}}) = W_\psi V_{\mathcal{S}} + b_\psi \in \mathbb{R}^C, \tag{40}$$

$$\ell_{\text{rerank}}(V_{\mathcal{S}}, y) = -r_{\psi,y}(V_{\mathcal{S}}) + \log \sum_{c \in \mathcal{S}(Z,y)} \exp r_{\psi,c}(V_{\mathcal{S}}), \tag{41}$$

and the label-aware training candidate set used by both variants is

$$\mathcal{S}(Z, y) = \begin{cases} \mathcal{S}(Z), & y \in \mathcal{S}(Z), \\ \big(\mathcal{S}(Z) \setminus \{c_{(K)}(Z)\}\big) \cup \{y\}, & y \notin \mathcal{S}(Z), \end{cases} \tag{42}$$

where $c_{(K)}(Z)$ is the lowest-scoring class in the natural top-K set, so $|\mathcal{S}(Z, y)| = K$ and the target is always present. For the SAE variant, $\mathcal{S}(Z, y)$ is also used to build the shared mask during training. The inference-time prediction is

$$\widehat{Y} = \arg\max_{c \in \mathcal{S}(Z)} h_c^{(\lambda)}(Z). \tag{43}$$

E.5 Full Statement of the Logit Fusion

Fusion as exponential tilting. Logit addition has an exact probabilistic reading. Define the global and reranking distributions over $\mathcal{S} = \mathcal{S}(Z)$,

$$p_G(c) = \frac{\exp s_c(Z)}{\sum_{k\in\mathcal{S}} \exp s_k(Z)}, \qquad p_R(c) = \frac{\exp r_c}{\sum_{k\in\mathcal{S}} \exp r_k}. \tag{44}$$

Proposition 3 (Global–local fusion as exponential tilting)**.** *The candidate-restricted distribution induced by Equation 17 satisfies*

$$p_\lambda(c) = \frac{p_G(c)\exp(\lambda r_c)}{\sum_{k\in\mathcal{S}} p_G(k)\exp(\lambda r_k)} \tag{45}$$

$$= \frac{p_G(c)p_R(c)^\lambda}{\sum_{k\in\mathcal{S}} p_G(k)p_R(k)^\lambda}. \tag{46}$$

Consequently, $\lambda = 1$ yields the normalized product of the global and reranking probabilities. Moreover, for every $y, c \in \mathcal{S}$,

$$\log\frac{p_\lambda(y)}{p_\lambda(c)} = \log\frac{p_G(y)}{p_G(c)} + \lambda\big(r_y - r_c\big). \tag{47}$$

Finally, letting $\Delta(\mathcal{S})$ denote the probability simplex over $\mathcal{S}$, p_λ is the unique solution of

$$p_\lambda = \arg\max_{q\in\Delta(\mathcal{S})} \left\{\lambda\,\mathbb{E}_{c\sim q}[r_c] - D_{\mathrm{KL}}(q\|p_G)\right\}. \tag{48}$$

Thus, fusion favors candidates supported by the reranking head while penalizing excessive departure from the frozen global distribution.

Proof. Applying the candidate-restricted softmax to Equation 17,

$$p_\lambda(c) = \frac{\exp\big(s_c(Z) + \lambda r_c\big)}{\sum_{k\in\mathcal{S}} \exp\big(s_k(Z) + \lambda r_k\big)} = \frac{p_G(c)\exp(\lambda r_c)}{\sum_{k\in\mathcal{S}} p_G(k)\exp(\lambda r_k)}, \tag{49}$$

since the global softmax normalizer is class-independent and cancels. This is Equation 45. Because $\exp(\lambda r_c)$ and $p_R(c)^\lambda$ differ by a positive class-independent factor, Equation 46 follows. Taking the ratio of Equation 45 at y and c and the logarithm gives Equation 47.

For the variational form, add a Lagrange multiplier η for $\sum_{c\in\mathcal{S}} q(c) = 1$. The derivative of the objective in Equation 48 with respect to $q(c)$ is

$$\lambda r_c - \log\frac{q(c)}{p_G(c)} - 1 + \eta. \tag{50}$$

Setting it to zero gives $q(c) \propto p_G(c)\exp(\lambda r_c)$, and normalizing recovers Equation 45. Uniqueness holds because $-D_{\mathrm{KL}}(q\,\|\,p_G)$ is strictly concave on the simplex whenever $p_G(c) > 0$, which a softmax guarantees. □

The identity in Equation 46 is algebraic and needs no independence or calibration of the two heads. Equation 47 says the reranker adds $\lambda(r_y - r_c)$ to the frozen global log-odds, so a strong local preference can overturn a weak global one, and a strong global preference resists an incorrect local adjustment.

At $\lambda = 1$, fusion is a normalized product of the two candidate-level distributions, with no independence or calibration assumption. The variational form gives the division of labor: the expected reranking score rewards candidates the local head prefers, and the KL term anchors the result to the frozen global distribution. A strong local preference can overturn a weak global one; a strong global preference resists an incorrect local adjustment. For the SAE variant, the anchor also retains full-representation evidence when the prevalence rule removes an atom that still carried useful local information.

F Details of Experiment Settings

Details of datasets. We evaluate our approach on eleven image-classification benchmarks: ImageNet (Deng et al., 2009), CUB-200 (Wah et al., 2011), DTD (Cimpoi et al., 2014), Flowers102 (Nilsback & Zisserman, 2008), StanfordCars (Krause et al., 2013), OxfordPets (Parkhi et al., 2012), CIFAR100 (Krizhevsky et al., 2009), FGVCAircraft (Maji et al., 2013), Food101 (Bossard et al., 2014), Caltech101 (Fei-Fei et al., 2004), and UCF101 (Soomro et al., 2012). This collection covers diverse recognition regimes. ImageNet, Caltech101, and CIFAR100 evaluate general object recognition across different dataset scales and image resolutions; CUB-200, Flowers102, StanfordCars, OxfordPets, FGVCAircraft, and Food101 emphasize fine-grained distinctions between visually similar categories; DTD focuses on material and texture recognition; and UCF101 extends the evaluation to human-action categories. Together, these benchmarks vary substantially in domain, visual granularity, number of classes, and intra-class variation, allowing us to examine whether the global-to-local mismatch and the resulting reranking behavior generalize beyond a single type of classification problem.

Details of global classifiers. We instantiate the frozen first-stage global classifier using three common adaptation strategies: a linear classifier, CoOp, and CoCoOp (Zhou et al., 2022b;a). The linear classifier, also referred to as a linear probe, learns a single classification layer on top of frozen CLIP image representations. It is the simplest and most fundamental form of supervised adaptation and therefore provides a direct test of whether candidate-restricted reranking is effective without a specialized prompt-learning mechanism. CoOp instead adapts CLIP by learning continuous context vectors in the text prompts used to construct the class prototypes, while keeping the pretrained encoders frozen. CoCoOp extends this formulation by conditioning the prompt on each input image, allowing the resulting class prototypes to adapt to individual examples. After each global classifier is trained, all of its parameters are frozen during reranker training.

Training details. All experiments are conducted on a single NVIDIA RTX 3090 GPU. Unless otherwise stated, the candidate-set size is $K = 15$, the class-prevalence threshold is $\rho = 0.75$, and the atom-activation threshold is $\delta = 0$. We train each reranking head using AdamW (Loshchilov & Hutter, 2019) with a batch size of 256, an initial learning rate of 5×10^{-3}, 50 warm-up steps, and a cosine learning-rate schedule. Training runs for 50 epochs by default; we use 25 epochs for experiments with CoCoOp on ImageNet due to their substantially higher computational cost. The global–local fusion coefficient is fixed to $\lambda = 1.0$ in every experiment and is not tuned for individual datasets or classifiers.

G More Experimental Results

Full results with CLIP–ResNet-50 and CLIP–ViT-B/32. Table 2 shows that candidate-restricted reranking improves the macro-average accuracy of all three global classifiers at 4, 8, and 16 shots, whether the reranker receives the full representation Z or the SAE residual $V_{\mathcal{S}}$. The gains persist as the number of training examples decreases. This consistency across classifier types and shot settings suggests that an additional candidate-local discrimination step can refine decisions made by the global classifier.

Table 3 shows that the findings extend to a transformer-based visual encoder. Candidate-restricted reranking consistently improves both Linear Probe and CoOp across the 4-, 8-, and 16-shot settings. Moreover, reranking on $(V_{\mathcal{S}})$ remains comparable to reranking on (Z), with their macro-averages differing by at most (0.06) percentage points. Together with the ResNet-50 results in Table 2, these findings indicate that the observed behavior is consistent across convolutional and transformer-based CLIP architectures.

Effect of the candidate-set size. The choice of K balances candidate coverage against the locality of the reranking problem. With a small K, the ground-truth class is more likely to be excluded from the retrieved set, limiting the maximum achievable reranking accuracy. Increasing K improves coverage, but also introduces less plausible classes into the local decision. Moreover, because an SAE atom is considered shared only when it is prevalent in every candidate class, the intersection of shared atoms becomes smaller as K increases. Consequently, fewer shared components are

Table 2: Main results on 11 image-classification datasets using CLIP–ResNet-50 in the 4-, 8-, and 16-shot, top-15 settings. Entries are top-1 accuracy Acc_h (%). Full-representation reranking applies the candidate-restricted objective to Z, whereas SAE-residual reranking uses V_S after removing candidate-shared concepts. **Bold** and underlined values denote the best and second-best results within each global-classifier and shot-setting block.

Global classifier	Reranker input	Dataset											Avg. (Acc_h)
		ImageNet	CUB-200	DTD	Flowers102	StanfordCars	OxfordPets	CIFAR100	FGVCAircraft	Food101	Caltech101	UCF101	
CLIP (zero-shot)		55.49	46.68	42.39	66.43	54.17	85.66	38.98	17.07	80.59	83.16	57.81	57.13
16-shot													
Linear Probe	None	56.14	65.97	66.70	92.69	69.99	76.80	50.96	35.03	74.32	91.19	74.42	68.56
	Z	**58.36**	**67.58**	**67.60**	92.73	**71.88**	**78.00**	52.33	**35.94**	76.21	**92.11**	**75.96**	**69.88**
	V_S	58.20	67.35	67.51	**92.79**	71.35	77.75	**52.58**	35.92	**76.25**	92.08	**75.96**	69.79
CoOp	None	60.30	66.33	64.64	91.51	72.00	89.63	53.89	32.99	82.66	91.99	76.48	71.13
	Z	**64.23**	**70.07**	**66.50**	92.74	75.24	**90.16**	55.49	35.66	**83.23**	**93.33**	**78.77**	**73.22**
	V_S	64.16	70.04	66.48	**92.81**	**75.25**	90.05	**55.55**	**35.77**	83.21	93.30	78.71	73.21
CoCoOp	None	60.93	57.13	59.41	80.94	64.58	89.57	50.11	20.12	82.58	91.41	72.67	66.31
	Z	**64.77**	**69.47**	**65.03**	89.55	**75.05**	90.16	54.91	31.07	83.22	**93.11**	**78.42**	**72.25**
	V_S	64.70	**69.47**	64.76	**89.61**	74.63	**90.17**	**55.26**	**31.26**	**83.24**	93.10	78.28	72.23
8-shot													
Linear Probe	None	49.79	57.02	60.44	90.94	61.14	68.19	43.83	28.75	67.96	87.73	69.30	62.28
	Z	**52.29**	**59.06**	**61.75**	91.13	**62.88**	**69.84**	45.25	**29.47**	**69.95**	**89.11**	**71.21**	**63.81**
	V_S	52.15	58.27	61.55	**91.23**	62.60	69.43	**45.40**	29.42	69.85	89.09	70.96	63.63
CoOp	None	59.01	60.65	60.65	90.32	67.03	88.78	50.54	28.44	82.05	90.99	72.54	68.27
	Z	**62.19**	64.06	**62.43**	91.89	**70.09**	**89.23**	52.44	30.07	**82.41**	**91.96**	**75.01**	70.16
	V_S	62.11	**64.09**	62.42	**91.96**	69.89	89.20	**52.64**	**30.22**	82.40	91.95	74.94	**70.17**
CoCoOp	None	58.78	53.89	53.79	78.84	62.06	89.07	48.07	17.46	82.17	90.58	70.34	64.10
	Z	**62.28**	63.99	**59.85**	87.49	**70.35**	89.45	51.97	25.35	**82.50**	92.16	74.49	**69.08**
	V_S	62.14	**64.10**	**59.85**	**87.68**	69.74	**89.46**	**52.26**	**25.46**	82.46	**92.17**	**74.53**	**69.08**
4-shot													
Linear Probe	None	41.79	46.24	52.48	84.88	49.69	56.26	35.73	23.53	57.98	83.50	62.83	54.08
	Z	**44.15**	**47.39**	**54.31**	85.60	**50.74**	**58.82**	37.79	**24.59**	**60.31**	85.48	**63.89**	**55.73**
	V_S	43.88	46.80	54.29	**85.61**	50.43	58.48	**37.96**	24.21	60.06	**85.54**	63.62	55.53
CoOp	None	58.80	55.65	54.70	84.99	61.81	87.02	47.69	23.89	80.29	89.78	68.83	64.86
	Z	**60.92**	**59.40**	**56.34**	87.13	**65.27**	**87.56**	**49.99**	**25.88**	**80.86**	**90.72**	**71.04**	**66.83**
	V_S	60.90	59.27	**56.34**	**87.16**	65.07	**87.56**	49.94	25.82	80.81	**90.72**	71.03	66.78
CoCoOp	None	58.18	52.32	50.07	72.86	59.33	88.39	46.72	10.84	82.16	89.49	66.42	61.53
	Z	**60.51**	**59.43**	55.21	80.60	**65.11**	**88.76**	49.29	17.87	**82.56**	**91.17**	**70.21**	65.52
	V_S	60.43	59.39	**55.30**	**80.70**	65.09	88.69	**49.36**	**18.21**	82.51	91.15	70.20	**65.55**

removed, and the candidate-restricted objective gradually approaches the broader global discrimination setting analyzed in Proposition 1.

Figure 7 shows that average reranking accuracy increases from 64.60% at $K = 5$ to 68.33% at $K = 15$, while further increases produce only marginal gains: 68.65% at $K = 20$ and 68.87% at $K = 30$. The dataset-level results exhibit the same general pattern, with performance typically stabilizing around $K = 15$. We therefore use $K = 15$ as the smallest candidate set near the performance plateau, providing a favorable balance among retrieval coverage, candidate-local specificity, and reranking cost.

Effect of the prevalence threshold and the effectiveness of the hybrid inference. The threshold ρ controls how conservatively an SAE atom is identified as shared: low values select and remove a broader set of atoms, whereas high values retain only concepts that occur consistently in every candidate class. As ρ increases, fewer atoms enter U_S, and the residual V_S approaches the complete representation Z. Accordingly, Figure 8 shows that average standalone reranking accuracy increases

Table 3: Results on 11 image-classification datasets using CLIP–ViT-B/32 in the 4-, 8-, and 16-shot, top-15 settings. Entries are top-1 accuracy Acc_h (%). Full-representation reranking applies the candidate-restricted objective to Z, whereas SAE-residual reranking uses V_S after removing candidate-shared concepts. **Bold** and underlined values denote the best and second-best results within each global-classifier and shot-setting block. CoCoOp is omitted because of its substantial training time.

Global classifier	Reranker input	Dataset											Avg. (Acc_h)
		ImageNet	CUB-200	DTD	Flowers102	StanfordCars	OxfordPets	CIFAR100	FGVCAircraft	Food101	Caltech101	UCF101	
CLIP (zero-shot)		59.32	53.04	44.63	66.73	58.86	87.46	61.69	18.96	84.15	90.58	61.11	62.41
16-shot													
Linear Probe	None	61.29	70.43	68.75	93.31	73.78	80.44	67.43	36.45	79.91	94.18	78.59	73.14
	Z	61.94	**71.04**	69.71	93.65	**74.85**	81.44	68.72	**37.02**	**81.27**	**94.80**	**79.38**	**73.98**
	V_S	**61.99**	70.96	**69.89**	**93.78**	74.42	**81.59**	**68.83**	36.76	81.02	94.79	79.19	73.93
CoOp	None	65.65	68.20	67.52	91.79	73.91	87.33	70.06	35.15	85.25	95.10	78.74	74.43
	Z	67.79	**73.21**	69.80	93.63	**77.82**	87.59	72.34	38.22	85.71	96.07	80.90	76.64
	V_S	**67.80**	72.98	**69.89**	**93.87**	77.56	**87.63**	**72.39**	**38.40**	**85.76**	**96.10**	**81.02**	**76.67**
8-shot													
Linear Probe	None	55.78	62.12	62.97	91.80	65.53	74.55	61.58	30.43	75.04	92.30	74.35	67.86
	Z	56.72	62.65	**64.04**	**92.39**	**66.99**	**75.87**	63.20	**31.03**	76.28	92.96	**75.59**	**68.88**
	V_S	**56.85**	**62.69**	63.97	92.36	66.56	75.60	**63.44**	30.93	**76.31**	**93.10**	75.40	68.84
CoOp	None	64.46	63.05	62.35	90.47	69.30	86.33	68.59	29.87	85.11	94.80	76.49	71.89
	Z	66.19	**67.81**	65.10	92.72	**73.09**	**86.75**	70.23	**32.29**	85.22	**95.57**	**78.98**	**74.00**
	V_S	**66.33**	67.78	**65.12**	**92.79**	72.99	86.72	**70.26**	32.28	**85.24**	95.46	78.80	73.98
4-shot													
Linear Probe	None	48.13	51.32	55.21	87.04	54.69	63.79	53.65	26.33	67.04	88.98	67.73	60.36
	Z	**49.22**	**51.91**	**56.50**	**87.76**	**55.36**	**65.51**	55.76	**26.54**	68.50	90.17	**68.78**	**61.46**
	V_S	49.07	51.51	56.47	87.61	55.21	65.38	**55.98**	26.46	**68.55**	**90.33**	**68.78**	61.40
CoOp	None	63.41	58.40	56.61	85.94	65.38	84.15	67.04	26.14	84.24	93.79	73.50	68.96
	Z	64.67	**62.65**	**60.37**	89.42	**68.63**	84.16	**68.83**	**28.72**	84.47	**94.55**	**76.39**	**71.17**
	V_S	**64.69**	**62.65**	60.36	**89.51**	68.44	**84.24**	68.51	28.70	**84.82**	94.47	76.36	71.16

monotonically from 65.84% at $\rho = 0$ to 68.87% at $\rho = 1$; under the strict criterion $\widehat{\pi}_{c,j} > \rho$, the latter selects no atoms and recovers full-Z reranking. In contrast, hybrid accuracy remains between 69.74% and 69.91% across all thresholds, showing that the final prediction is insensitive to how aggressively shared evidence is withheld from the local branch. We use $\rho = 0.75$ because it provides a conservative and interpretable definition of sharedness while retaining a nontrivial intervention: its hybrid accuracy of 69.79% is within 0.12 points of the best result and within 0.09 points of the no-removal endpoint. Thus, the selected threshold preserves the explicit concept-level intervention without relying on accuracy-based tuning.

Classwise prevalence–magnitude agreement. Figure 9 expands the aggregate analysis in Section 2.1 by reporting the complete distribution for every dataset. Across all eleven datasets, the distributions remain strongly positive, with Fisher-averaged correlations ranging from 0.67 to 0.85. Most classes exhibit high prevalence–magnitude agreement, showing that the dataset-level result is not driven by a small subset of categories. This consistency corroborates our central observation: SAE atoms that are prevalent across a class's probable candidates also tend to be expressed strongly, supporting prevalence as an interpretable indicator of prominent candidate-shared structure.

Classwise score deviations. Figure 10 expands the aggregate analysis by reporting $|b_U(y, c^\star)|$ for each ground-truth class. Across all eleven datasets, the stage-1 classifier exhibits nonzero and occasionally large deviations, confirming that concept-level sharedness does not imply exact score-level cancellation. Candidate-restricted reranking on Z consistently reduces these deviations across the label space, showing that the aggregate reduction is not driven by a small subset of classes. This

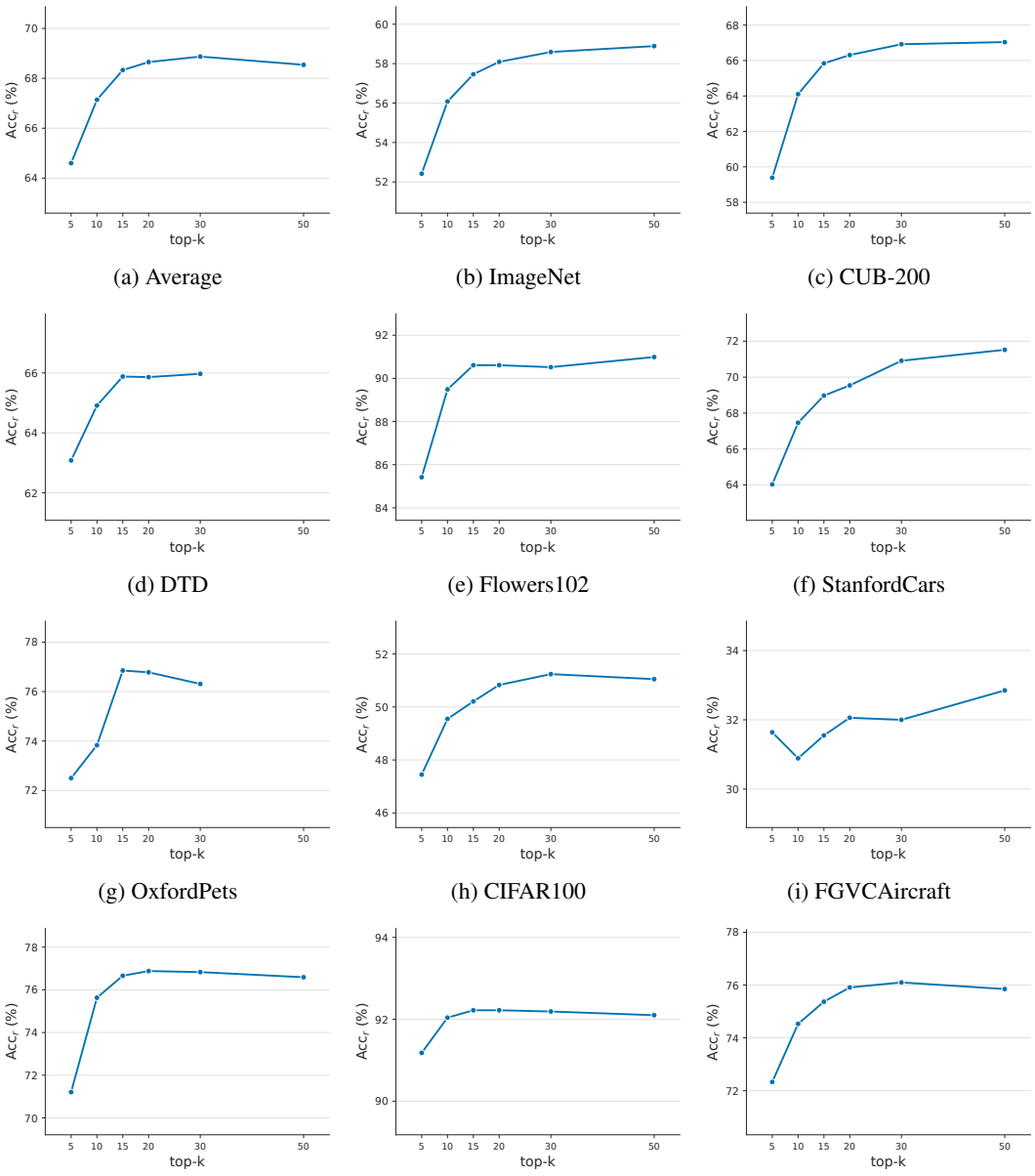
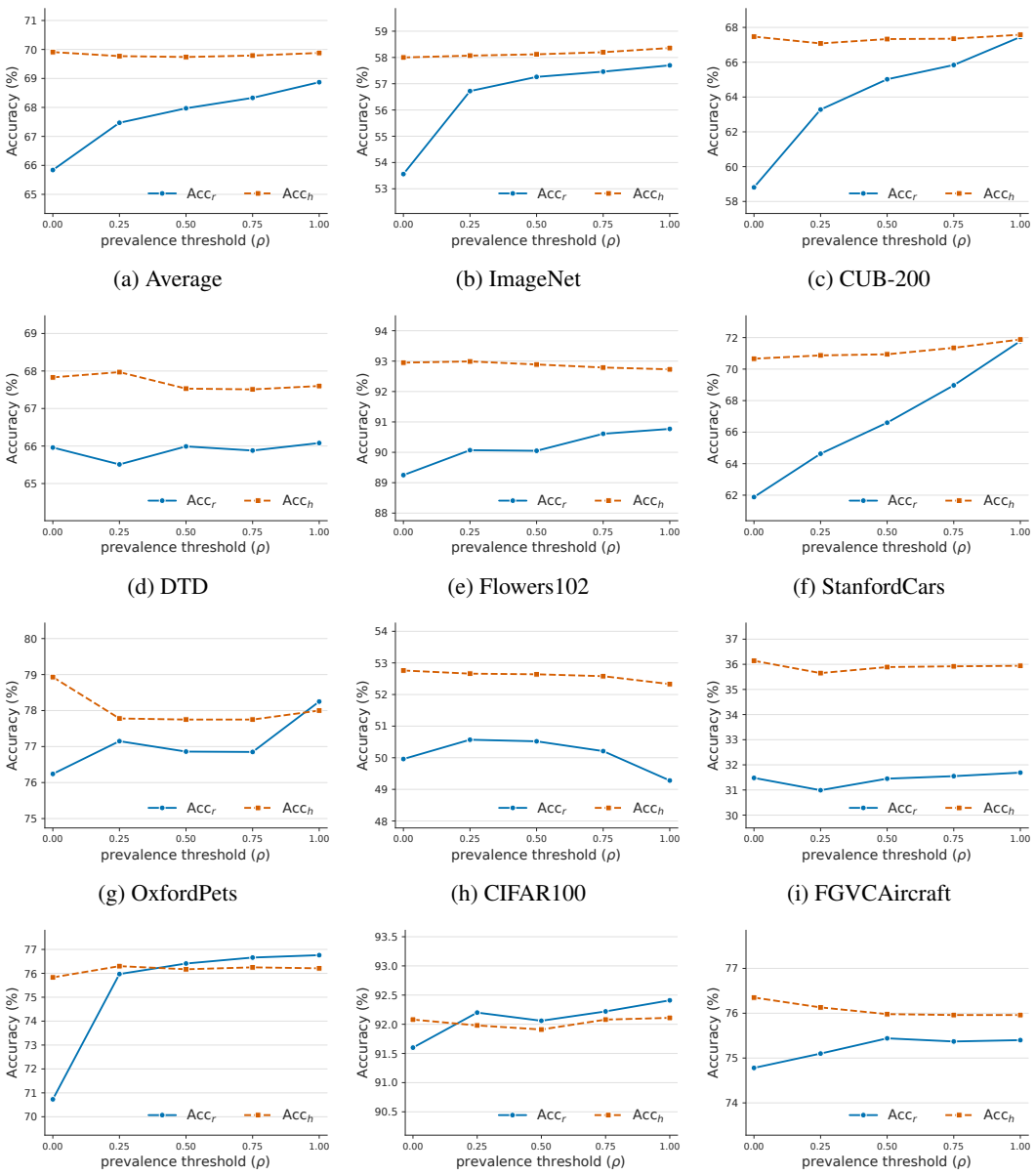


Figure 7: Effect of the candidate-set size K on the accuracy of the trained residual reranker in the 16-shot setting. The first panel reports the macro-average, followed by the results on the eleven datasets. Accuracy improves rapidly between $K = 5$ and $K = 15$, after which the gains become marginal. The aggregate result at $K = 50$ averages the datasets for which results are available.

classwise consistency further supports the imperfect-cancellation mechanism in Proposition 2 and the implicit adaptation learned by the full-Z reranker. The residual reranker is omitted because U_S is excluded from its input, making its accessible contribution zero by construction.

Qualitative analysis. Figures 11 to 15 present successful corrections and representative failure modes. These examples illustrate that the frozen global classifier can assign comparable scores to several plausible top-(K) classes, providing insufficient separation for the correct prediction. In successful cases, candidate-shared evidence is removed and the candidate-restricted reranker supplies sufficient relative evidence to resolve this confusion, allowing the hybrid score to recover the ground-



(a) Average (b) ImageNet (c) CUB-200

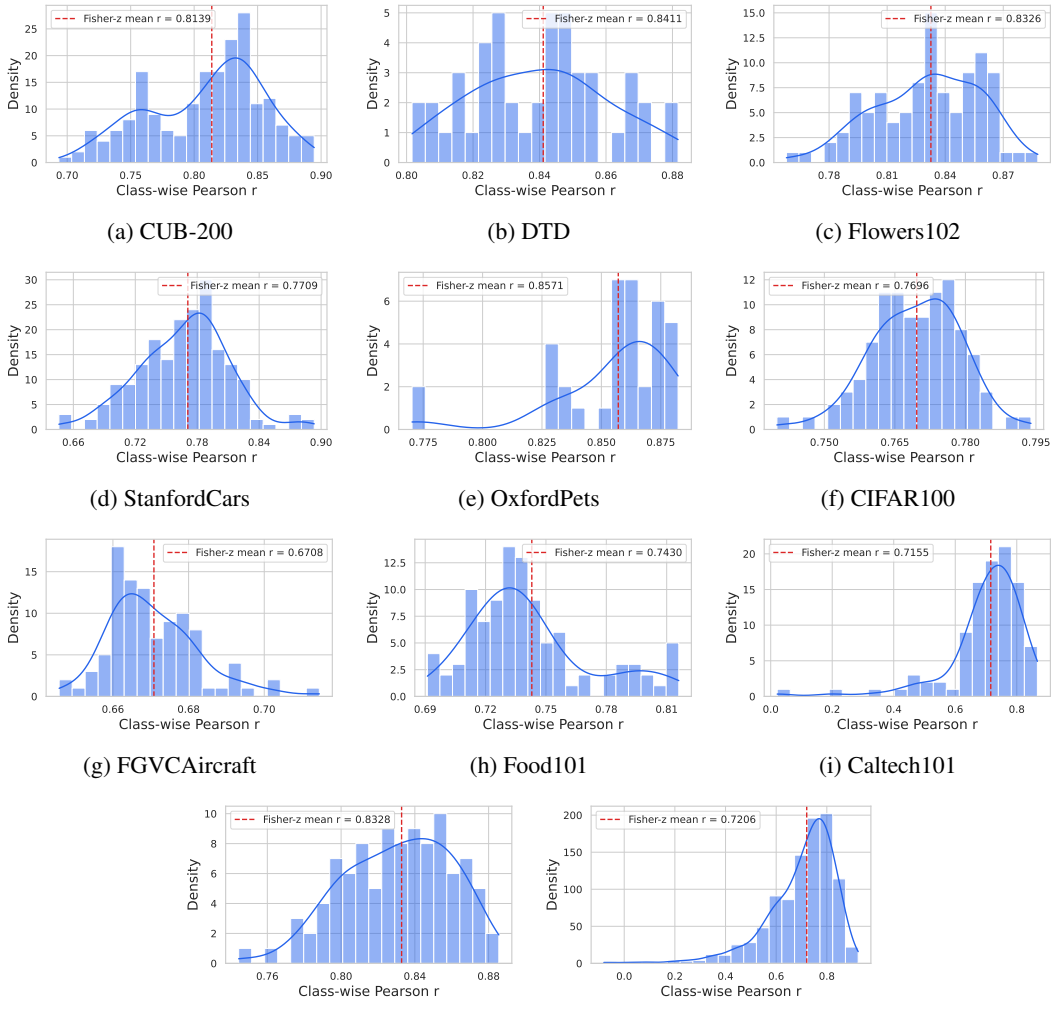
(d) DTD (e) Flowers102 (f) StanfordCars

(g) OxfordPets (h) CIFAR100 (i) FGVCAircraft

(j) Food101 (k) Caltech101 (l) UCF101

Figure 8: Effect of the class-level prevalence threshold ρ on standalone reranking accuracy Acc_r and hybrid accuracy Acc_h in the 16-shot, top-15 setting. The first panel reports the macro-average, followed by the eleven dataset-specific results. Increasing ρ makes the shared-atom criterion more selective and removes fewer components. Standalone reranking accuracy consequently increases, whereas hybrid accuracy remains stable across the full threshold range.

truth class. A recurring failure mode occurs when the reranker alone ranks the ground-truth class correctly, but its contribution is insufficient to overcome the original Stage-1 margin. Such cases reflect a scale mismatch between the global and local scores rather than a failure of candidate-local ordering. We keep the fusion coefficient (λ=1) throughout our experiments; calibrating or adaptively selecting (λ) may allow more successful local corrections to propagate to the final prediction.



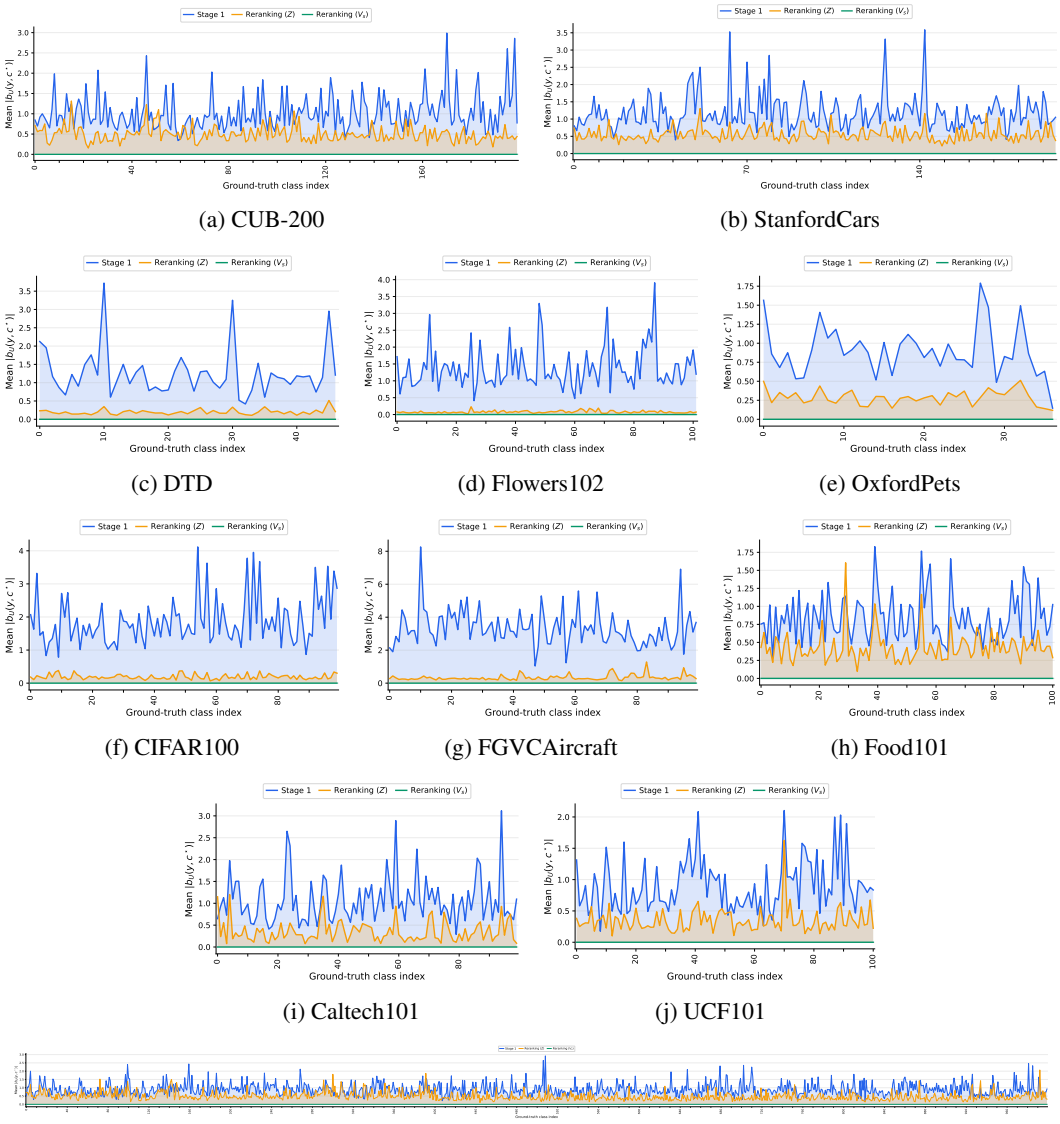
(a) CUB-200 (b) DTD (c) Flowers102

(d) StanfordCars (e) OxfordPets (f) CIFAR100

(g) FGVCAircraft (h) Food101 (i) Caltech101

(j) UCF101 (k) ImageNet

Figure 9: Classwise distributions of prevalence–magnitude agreement across all eleven datasets. Each panel shows the distribution of the Pearson correlation r_c between prevalence-based sharedness $Q^P_{\mathcal{S}_c,j}$ and magnitude-based sharedness $Q^M_{\mathcal{S}_c,j}$ across valid SAE atoms for each target class c. The red dashed line denotes the class-balanced mean computed on the Fisher z scale. The distributions are concentrated at strongly positive correlations, showing that the aggregate agreement reported in the main text is broadly consistent across classes.



(a) CUB-200 (b) StanfordCars

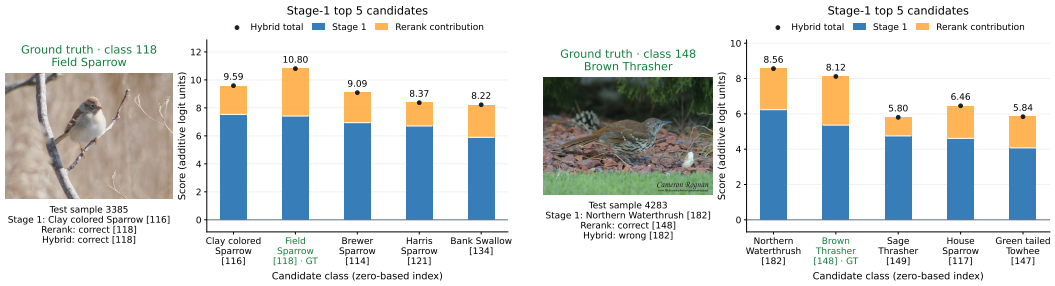
(c) DTD (d) Flowers102 (e) OxfordPets

(f) CIFAR100 (g) FGVCAircraft (h) Food101

(i) Caltech101 (j) UCF101

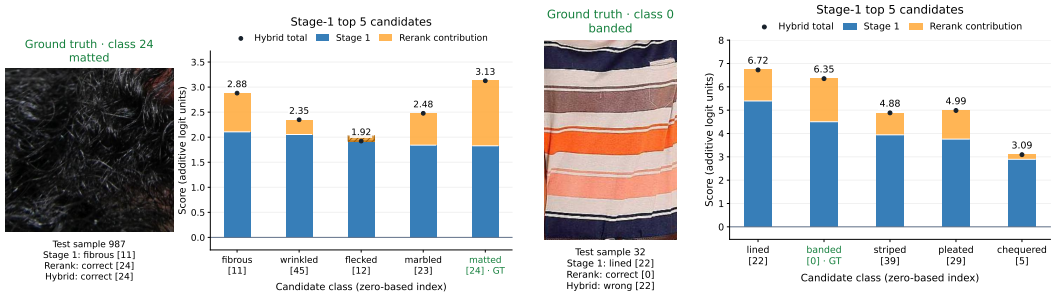
(k) ImageNet

Figure 10: Classwise magnitude of the uncancelled candidate-shared contribution $|b_U(y, c^\star)|$ across all datasets. For each example, y is the ground-truth class and $c^\star$ is the strongest competing candidate retrieved by the frozen stage-1 classifier. Each point reports the mean over examples from the corresponding ground-truth class. Candidate-restricted reranking on Z consistently reduces the contribution relative to the stage-1 classifier across the label space. For reranking on V_S, the accessible contribution of U_S is zero by construction.



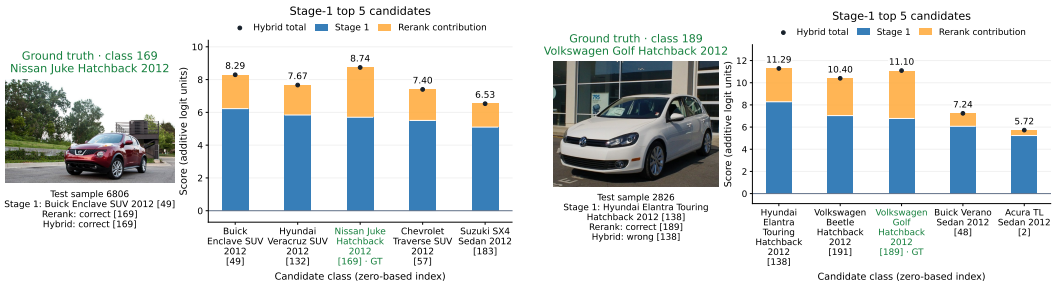
(a) Positive example: successful hybrid correction. (b) Negative example: insufficient hybrid correction.

Figure 11: Qualitative reranking examples on CUB-200. The class order follows the Stage-1 predictions.



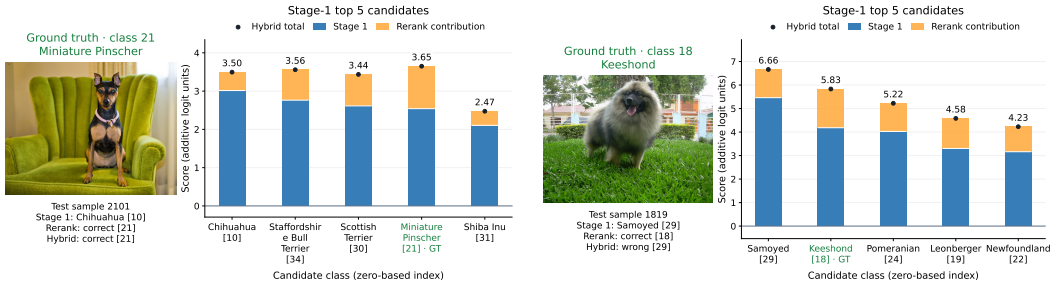
(a) Positive example: successful hybrid correction. (b) Negative example: insufficient hybrid correction.

Figure 12: Qualitative reranking examples on DTD. The class order follows the Stage-1 predictions.



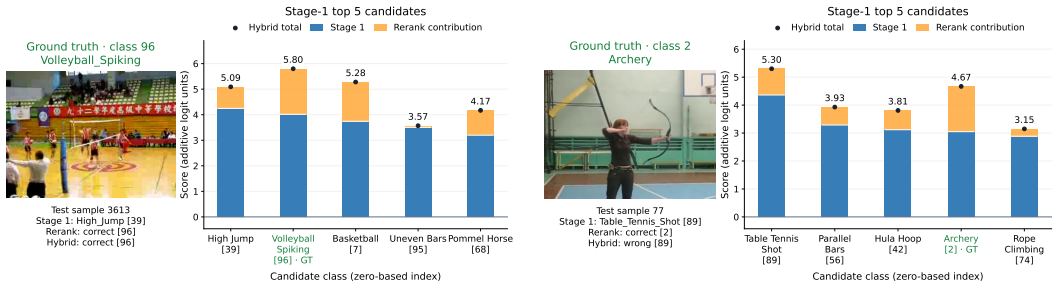
(a) Positive example: successful hybrid correction. (b) Negative example: insufficient hybrid correction.

Figure 13: Qualitative reranking examples on StanfordCars. The class order follows the Stage-1 predictions.



(a) Positive example: successful hybrid correction. (b) Negative example: insufficient hybrid correction.

Figure 14: Qualitative reranking examples on OxfordPets. The class order follows the Stage-1 predictions.



(a) Positive example: successful hybrid correction.

(b) Negative example: insufficient hybrid correction.

Figure 15: Qualitative reranking examples on UCF101. The class order follows the Stage-1 predictions.

H Limitations and Broader Impact

Our work has limitations. First, all candidate-local results are conditioned on successful retrieval, so no reranker can recover a class the first stage omits from the top-K set, similar to Chen et al. (2023). Second, we select which atoms to remove with a simple frequency rule: an atom counts as shared if it fires in more than a fixed fraction of the training images of every candidate class. What Assumption 1 actually requires is stronger: that removing those atoms costs no information for choosing among the candidates. Frequency is evidence for that condition and not proof of it, so the atoms we remove satisfy the assumption only approximately, and a case where a frequent atom still carries discriminative signal is possible. A natural next step is to apply candidate conditioning where the candidate set is given rather than retrieved, such as multiple-choice question answering in language models.

Our findings establish candidate conditioning as a general framework for understanding how the relevance of representation components changes with the prediction context. This perspective can be applied to systems that first retrieve or generate a set of plausible candidates and then make a finer decision, including hierarchical and fine-grained classification, retrieval reranking, and candidate-based reasoning tasks such as multiple-choice question answering. By distinguishing evidence that supports broad recognition from evidence that separates the remaining alternatives, the framework can help diagnose errors, analyze objective mismatch between global and local decisions, and guide context-dependent feature selection or representation simplification. Our candidate-restricted reranking experiments provide one concrete demonstration of this broader idea, showing that locally discriminative performance can be improved while a large candidate-shared component is suppressed. Thus, the main potential impact is a general way to design and analyze the local decision stage of coarse-to-fine prediction systems across modalities and tasks.