Active Learning as Nullspace Regulation: A Spectral Representation Perspective

Luhang Shen **Jingwei Hu** **Na Ying** **Chunsheng Guo**
College of Communication Engineering, Hangzhou Dianzi University
Hangzhou 310018, China
gcs@hdu.edu.cn

Abstract

Most active learning (AL) methods estimate sample value through predictive uncertainty, diversity, or distributional coverage. We formulate AL from a spectral-geometric perspective: the leading singular directions of the current labeled-feature matrix define a modeled subspace, while its orthogonal complement provides an operational residual subspace for directions weakly represented by current supervision. We introduce a unified nullspace-guided acquisition framework with three coupled components: adaptive spectral subspace identification, cross-cycle tracking of relative residual change, and coverage-aware scoring in residual coordinates. The geometric score combines relative residual change with local residual sparsity; rank-normalized fusion integrates it with predictive entropy and learned loss prediction, and greedy residual-space diversity constructs each batch. Experiments on visual benchmarks show competitive or improved sample efficiency over strong baselines. Component ablations quantify the contributions of geometric and prediction-level cues. Further analyses reveal anisotropic residual contraction and substantial cross-backbone differences in acquisition rankings, suggesting that sample value depends on the current representation state rather than being purely data-intrinsic.

1 Introduction

Deep neural networks have achieved remarkable success in visual recognition, but their performance typically depends on large amounts of labeled data [1]. Active learning (AL) aims to reduce annotation cost by iteratively selecting informative samples from a large unlabeled pool [2, 3]. Despite substantial progress, a unified explanation of why a queried sample is informative for a given model remains incomplete [4].

Most existing AL methods characterize informativeness through empirical criteria such as prediction uncertainty [5, 6], diversity [7, 8], or hybrid criteria [9]. These criteria operate on model predictions or learned features. Uncertainty-based methods select samples with ambiguous predictions, diversity-based methods encourage dispersion among selected samples, and representativeness-based methods improve coverage of observed data patterns. However, such criteria do not directly model which directions in representation space are weakly supported by the current labeled set. As a result, the connection between sample selection and representation improvement often remains indirect: informativeness is usually associated with prediction ambiguity or distributional coverage, rather than with a model-aware estimate of directions that are insufficiently represented by current supervision, as shown in Figure 1.

In this work, we revisit active learning from a spectral-geometric perspective. At each acquisition cycle, the labeled data induce a feature subspace corresponding to directions most strongly supported by current supervision. We use its orthogonal complement as an operational proxy for nullspace

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

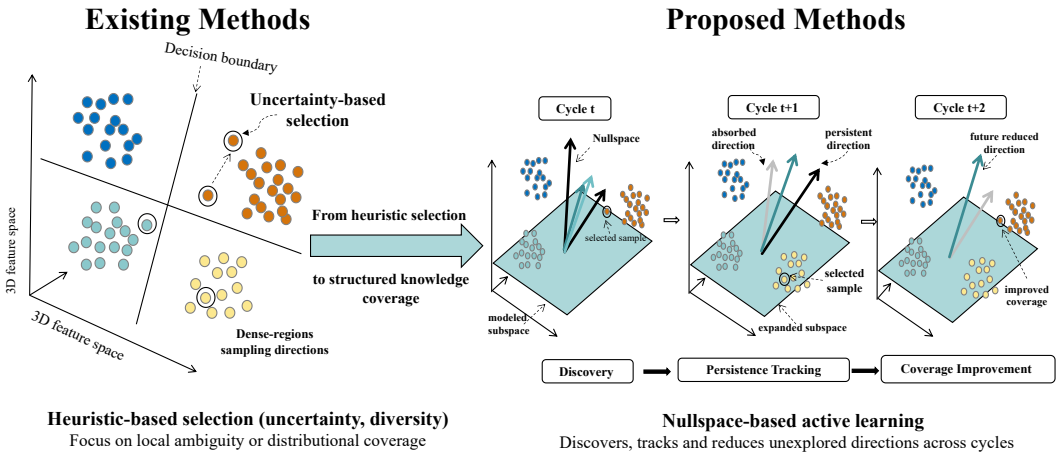


Figure 1: Conceptual comparison of acquisition views (schematic). Left: prediction ambiguity or distributional coverage. Right: cycle-specific modeled directions and operational residual directions, with persistence tracking and coverage-aware acquisition. Continued training updates the representation and the subspace decomposition at each cycle.

residual directions, i.e., directions comparatively weakly supported by the current labeled representation [10–12]. This decomposition connects acquisition to the evolving coverage of residual directions outside the current labeled subspace.

This formulation leads to three practical challenges. First, the modeled/nullspace boundary should adapt to the evolving labeled-feature spectrum rather than rely on a fixed dimension or threshold. Second, because the nullspace changes as new labels are acquired, absolute residual magnitudes are not directly comparable across cycles. Third, batch acquisition should avoid repeatedly selecting samples along similar residual directions and instead improve coverage of under-explored residual modes. To address these issues, we propose a nullspace-guided hybrid criterion. It combines adaptive spectral boundary estimation and cross-cycle residual persistence with residual-density coverage, predictive entropy, and learned loss prediction. Fixed-weight score fusion is followed by greedy residual-space diversity selection. The labeled-feature spectrum updates the modeled boundary at each cycle, allowing acquisition to track changes in residual coverage.

We evaluate the framework on SVHN, CIFAR-100, and Caltech-101 [13–15], with convolutional networks [16], vision transformers [17], and foundation models [18]. The results show competitive or improved sample efficiency, and a long-tailed TinyImageNet stress test extends the evaluation to imbalanced class coverage. Our analyses track anisotropic residual contraction, persistent long-tail directions, and cross-backbone acquisition agreement [19]. These measurements give a spectral-geometric account of the budget and representation dependence observed in prior AL studies [20, 4]. Our contributions are summarized as follows:

- We formulate active learning from a spectral-geometric perspective, where the orthogonal complement of the labeled-data subspace is used as an operational proxy for nullspace residual directions that are weakly supported by the current labeled representation.
- We propose a nullspace-guided hybrid acquisition criterion combining spectral residual persistence and coverage with entropy, predicted loss, and residual-space batch diversity, and quantify the contributions of these components.
- We document anisotropic residual contraction, persistent long-tail directions, and representation-dependent acquisition rankings, alongside competitive or improved sample efficiency on three visual benchmarks.

2 Related Work

Active learning. In pool-based AL, acquisition criteria include uncertainty, diversity, representativeness or coverage, and combinations of these principles. Traditional AL primarily relies on

uncertainty-based criteria[6] or diversity-based methods[21]. These two families have been extensively reviewed in deep active learning and continue to serve as strong baselines in recent large-scale settings [22, 23]. While empirically successful, these heuristics usually define informativeness through model predictions or data distribution coverage [24, 25]. They do not explicitly model the residual directions that are weakly supported by the current labeled representation. Different from these empirical selection paradigms, our work studies AL through a nullspace-based geometric proxy, where sample value is estimated by whether a sample exposes persistent or under-covered residual directions in representation space.

Spectral Structure and Representation Geometry. Analyzing deep representations through spectral bias [10] and low-dimensional subspaces has provided useful insights into generalization and feature organization [26–29]. Prior works have used low-rank structure, feature collapse, intrinsic dimension, or subspace geometry to study representation organization, redundancy, and generalization [30, 31]. However, such geometric quantities are often used as static descriptors or one-shot scoring signals. Our approach differs by embedding the spectral decomposition into the iterative AL loop. The modeled subspace and its orthogonal complement are updated as the labeled set evolves, allowing us to track how the residual nullspace representation changes across acquisition cycles.

Feature-based acquisition in pool-based AL. Feature-based acquisition measures diversity via embedding distances or coverage through core-set objectives [7]. TypiClust selects typical examples within feature clusters and studies how acquisition strategies vary with the labeling budget [20]. Related work further combines uncertainty, diversity, gradient embeddings, or submodular objectives for scalable batch acquisition [32]. These methods improve over purely prediction-level heuristics by exploiting feature-space information. However, they often treat feature statistics as fixed cues for coverage within the current distribution. In contrast, we focus on the evolution of the orthogonal complement of the labeled-data subspace. Rather than only seeking representativeness within the learned feature distribution, our framework prioritizes samples that are estimated to reveal persistent and under-covered nullspace residual directions.

Learning dynamics and data value. The value of data is closely related to the model's training dynamics and current state[33]. Existing studies on data influence, memorization, data valuation, and training dynamics have shown that samples contribute differently during learning [34–38]. Recent active learning methods also exploit temporal prediction changes or training dynamics to guide acquisition [39, 40]. However, most data-value studies focus on supervised learning with fixed datasets, while active learning involves an iterative feedback loop between sample acquisition, supervision, and representation update. Our work brings a dynamic geometric view to this setting. By tracking how the labeled-data subspace and its orthogonal complement evolve across cycles, we provide an empirical geometric account of how representation coverage changes and how sample utility shifts during progressive acquisition.

3 Method

We formulate a nullspace-guided hybrid acquisition criterion that combines spectral residual geometry, predictive entropy, and learned loss prediction, followed by residual-space batch diversity. Nullspace regulation updates the modeled subspace and its orthogonal complement from the labeled-feature spectrum at each acquisition cycle. The framework consists of three coupled components: (i) discrete spectral regulation for estimating the modeled boundary; (ii) continuity adjustment for tracking residual persistence across cycles; and (iii) nullspace-oriented sampling for selecting persistent and under-covered residual directions, fused with the two prediction-level scores before greedy batch selection. An overview is shown in Figure 2.

3.1 Discrete Spectral Regulation

Labeled-data subspace and nullspace residual. Let $f_{\theta_t}(\cdot)$ denote the model at active learning cycle t, and let $h_t(x) = f_{\theta_t}(x) \in \mathbb{R}^d$ be the feature embedding of sample x. Given the labeled set $\mathcal{D}_L^{(t)}$, we construct the centered labeled feature matrix

$$H_L^{(t)} = [h_t(x_1), \ldots, h_t(x_n)]^\top, \qquad \widetilde{H}_L^{(t)} = H_L^{(t)} - \mathbf{1}\mu_t^\top, \quad \mu_t = \frac{1}{n}\sum_{i=1}^{n} h_t(x_i). \tag{1}$$

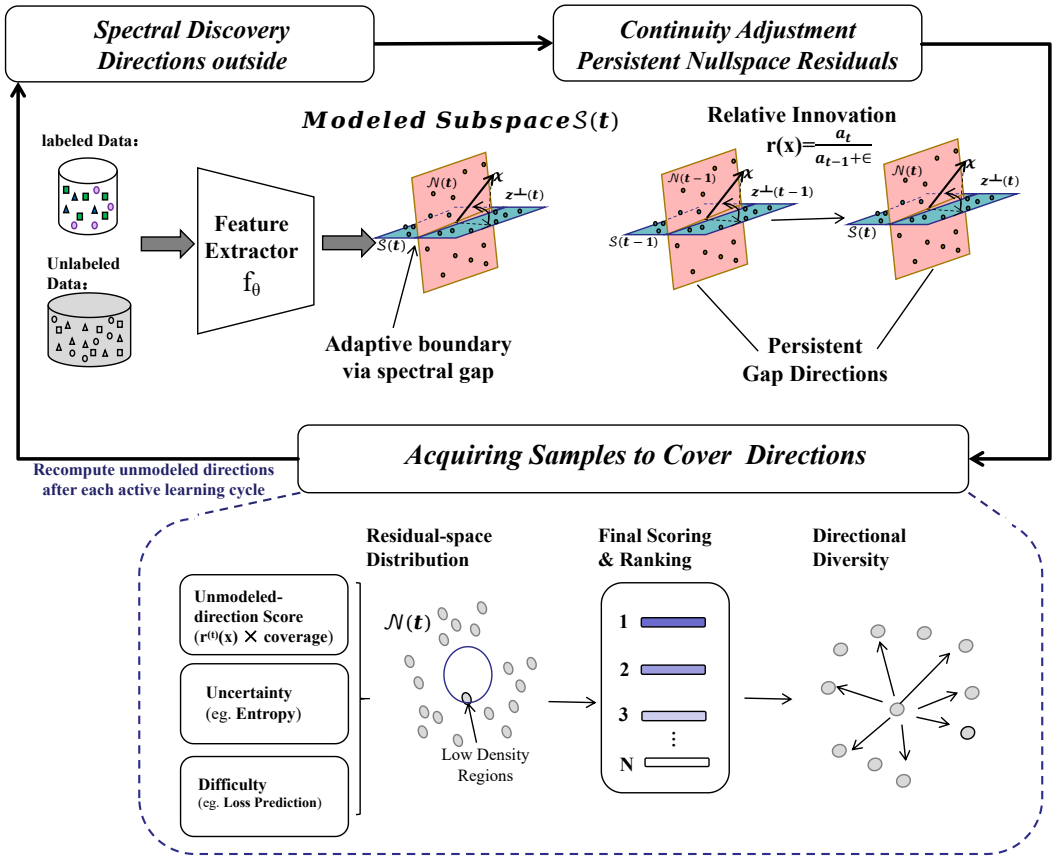


Figure 2: Complete hybrid acquisition pipeline. Residual magnitudes are cached under their own cycle-specific representations. Relative residual change and residual sparsity form the geometry score, which is rank-normalized and fused with entropy and predicted loss. Greedy residual-space diversity constructs the batch; annotation and continued training trigger recomputation.

We compute its singular value decomposition:

$$\widetilde{H}_L^{(t)} = U_t \Sigma_t V_t^\top, \qquad \Sigma_t = \mathrm{diag}(\sigma_1, \ldots, \sigma_r), \quad \sigma_1 \geq \cdots \geq \sigma_r. \tag{2}$$

Here $r = \min(n, d)$ is the number of singular values returned by the economy-size SVD, including any zero values; k_t is the selected modeled dimension, not r. Given k_t, the leading right singular vectors define the modeled subspace

$$\mathcal{S}^{(t)} = \mathrm{span}(V_{t,1:k_t}), \tag{3}$$

which represents the feature directions most strongly supported by the current labeled set. Its orthogonal complement defines the operational nullspace

$$\mathcal{N}^{(t)} = \left(\mathcal{S}^{(t)}\right)^{\perp}. \tag{4}$$

Thus, directions in $\mathcal{N}^{(t)}$ are treated as residual directions weakly supported by the current labeled-feature spectrum. For any sample x, we define the nullspace projection and residual representation as

$$P_{\mathrm{null}}^{(t)} = I - V_{t,1:k_t} V_{t,1:k_t}^\top, \qquad z_\perp^{(t)}(x) = P_{\mathrm{null}}^{(t)} h_t(x). \tag{5}$$

All subsequent persistence, coverage, and diversity terms are computed in this residual representation. Additional geometric interpretations are provided in Appendix A.1–A.3.

Adaptive boundary via spectral gap. Instead of fixing k_t manually, we estimate it from the singular spectrum at each acquisition cycle. Specifically, we normalize the singular values by

$$\bar{\sigma}_i = \frac{\sigma_i}{\max_j \sigma_j + \epsilon}, \tag{6}$$

and select the largest spectral drop:

$$k_t = \arg \max_{i \in \{1,\dots,r-1\}} (\bar{\sigma}_i - \bar{\sigma}_{i+1}) . \tag{7}$$

The resulting $V_{t,1:k_t}$ defines $\mathcal{S}^{(t)}$, while its orthogonal complement defines $\mathcal{N}^{(t)}$. This adaptive rule allows the modeled/nullspace split to evolve with the labeled-feature spectrum; its interpretation is discussed in Appendix A.4.

3.2 Continuity Adjustment for Nullspace Tracking

As active learning proceeds, newly acquired labels may expand or rotate the modeled subspace, causing the operational nullspace to change across cycles. Therefore, a large residual in one cycle does not necessarily imply persistent sample value in the next cycle. We track relative changes in residual magnitude after the labeled representation has been updated.

For an unlabeled sample $x \in \mathcal{D}_U^{(t)}$, the current residual magnitude is

$$a_t(x) = \left\| P_{\text{null}}^{(t)} h_t(x) \right\|_2 . \tag{8}$$

The previous-cycle reference is computed before the model update and cached as

$$a_{t-1}(x) = \left\| P_{\text{null}}^{(t-1)} h_{t-1}(x) \right\|_2 . \tag{9}$$

The cached scalar $a_{t-1}(x)$ and the current magnitude $a_t(x)$ summarize residual responses under their respective cycle-specific representations and decompositions.

We define the relative innovation score as

$$r^{(t)}(x) = \frac{a_t(x)}{a_{t-1}(x) + \epsilon}. \tag{10}$$

The ratio $r^{(t)}(x)$ measures relative residual change. When the previous residual is large compared with ϵ, values above one indicate an increase, values near one indicate comparable magnitudes, and small values indicate a decrease. A large ratio does not by itself imply a large absolute residual. In practice, residual-derived scores are rank-normalized over the remaining unlabeled pool to reduce sensitivity to feature-scale changes and extreme values. For the first acquisition cycle, where no previous residual cache is available, we set $r^{(t)}(x) = 1$ and rely on the current residual coverage score. Further discussion is provided in Appendix A.7.

3.3 Sampling Operation for Nullspace Coverage

We combine the relative residual-change signal with local sparsity in the current residual representation to select samples from under-explored regions.

Nullspace geometry score. For each $x \in \mathcal{D}_U^{(t)}$, we estimate the density of its residual representation by

$$\rho^{(t)}(x) = \sum_{x' \in \mathcal{D}_U^{(t)}} \exp\left(- \frac{\left\| z_\perp^{(t)}(x) - z_\perp^{(t)}(x') \right\|_2^2}{\sigma^2 + \epsilon} \right) . \tag{11}$$

The normalized density and sparsity-aware coverage score are

$$p^{(t)}(x) = \frac{\rho^{(t)}(x)}{\sum_{x' \in \mathcal{D}_U^{(t)}} \rho^{(t)}(x') + \epsilon}, \qquad c^{(t)}(x) = -\log\Big(p^{(t)}(x) + \epsilon \Big) . \tag{12}$$

A larger $c^{(t)}(x)$ indicates that the sample lies in a sparser residual region. Combining it with the relative residual-change score gives

$$s_{\text{um}}^{(t)}(x) = r^{(t)}(x) c^{(t)}(x), \tag{13}$$

which favors samples with larger relative residual change in under-covered regions.

Let $u_t(x)$ be classifier logits and $p_c(x) = \mathrm{softmax}(u_t(x))_c$ the predictive class probability. The auxiliary module g_t takes intermediate backbone features $\phi_t(x)$ and predicts a scalar classification loss, $s_{\mathrm{pl}}(x) = g_t(\phi_t(x))$. It is trained on labeled per-sample classification losses using the pairwise loss-prediction objective [41]. The complementary entropy score is:

$$s_{\mathrm{ent}}(x) = -\sum_{c=1}^{C} p_c(x) \log p_c(x), \qquad p_c(x) = \mathrm{softmax}(u_t(x))_c. \tag{14}$$

Each component is normalized independently over the current unlabeled pool of size M: $\widehat{s}_j(x) = (\mathrm{rank}_{\uparrow}(s_j(x)) - 1)/(M-1)$ for $M > 1$, where ascending sorting positions run from 1 to M. Larger values have higher priority. Sorting positions follow the implementation ordering, without averaging tied ranks; for a singleton pool, the normalized score is 1. The final acquisition score is

$$s^{(t)}(x) = \alpha \widehat{s}^{(t)}_{\mathrm{um}}(x) + \beta \widehat{s}_{\mathrm{ent}}(x) + \gamma \widehat{s}_{\mathrm{pl}}(x), \tag{15}$$

The fusion weights $(\alpha, \beta, \gamma) = (0.60, 0.25, 0.15)$ are fixed across datasets and acquisition rounds. Section 4 evaluates the contributions of these cues and the batch diversity term.

Diversity-aware batch selection. Selecting the top-B samples by $s^{(t)}(x)$ may produce redundant residual directions. We therefore perform greedy batch selection in the nullspace residual space. The first sample is

$$x_1 = \arg\max_{x \in \mathcal{D}_U^{(t)}} s^{(t)}(x). \tag{16}$$

Given the current batch $\mathcal{Q}$, the normalized distance from a candidate to the selected set is

$$d^{(t)}_{\mathcal{Q}}(x) = \frac{\min_{x' \in \mathcal{Q}} \left\| z^{(t)}_{\perp}(x) - z^{(t)}_{\perp}(x') \right\|_2}{\max\limits_{u \in \mathcal{D}_U^{(t)} \setminus \mathcal{Q}} \min\limits_{x' \in \mathcal{Q}} \left\| z^{(t)}_{\perp}(u) - z^{(t)}_{\perp}(x') \right\|_2 + \epsilon}. \tag{17}$$

Each subsequent sample is selected by

$$x^{\star} = \arg\max_{x \in \mathcal{D}_U^{(t)} \setminus \mathcal{Q}} \left[s^{(t)}(x) + \lambda_t d^{(t)}_{\mathcal{Q}}(x) \right], \qquad \lambda_t = \lambda_{\mathrm{div}} = 0.7. \tag{18}$$

The coefficient remains fixed across acquisition cycles, balancing the fused acquisition score and residual-space batch diversity. The greedy coverage interpretation is discussed in Appendix A.9.

4 Experiments

We evaluate acquisition accuracy across labeling budgets, residual-space evolution, cross-backbone acquisition agreement, and the contributions of individual components.

Datasets and backbones. We conduct experiments on SVHN, CIFAR-100, and Caltech-101 [13–15]. SVHN contains relatively structured digit images, CIFAR-100 contains fine-grained object categories, and Caltech-101 exhibits stronger category imbalance and heterogeneous visual structures. For the main comparison, we use Wide ResNet-18 for SVHN, ResNet-18 for CIFAR-100, and ImageNet-pretrained ResNet-50 for Caltech-101. For model-dependence analysis, we additionally evaluate ResNet-18, ViT, and CLIP-based representations.

Active learning protocol and baselines. All methods start from the same initial labeled set and select a fixed number of samples from the unlabeled pool at each cycle. After each acquisition round, training continues on the updated labeled set from the previous cycle's model parameters. The optimizers and learning-rate schedulers are reinitialized at each cycle. We compare with representative active learning methods, including Random, Coreset [7], Uncertainty (entropy sampling), LLoss [41], SCAL [42], BADGE [8], BALD [6], CEC [43], HSSAL [44], and TypiClust [20]. Unless otherwise specified, results are averaged over three random seeds with the same initial-budget protocol. Implementation details are provided in Appendix B.1.

Main Results. Table 1 reports accuracy averaged over three seeds; standard deviations for Ours and TypiClust are given in Appendix B.6. Ours has higher mean accuracy than TypiClust at every

Table 1: Main active learning results on three benchmarks. Accuracy (%) is reported as the mean over three seeds under different labeling budgets. Standard deviations for Ours and TypiClust are reported in Appendix B.6. For CIFAR-100, B1–B5 correspond to 1k, 2k, 3k, 4k, and 5k labels. For SVHN and Caltech-101, B1–B6 correspond to 0.5k, 1k, 1.5k, 2k, 2.5k, and 3k labels.

Method	CIFAR-100					SVHN						Caltech-101					
	B1	B2	B3	B4	B5	B1	B2	B3	B4	B5	B6	B1	B2	B3	B4	B5	B6
Random	13.4	18.4	23.9	28.9	34.1	31.1	57.9	73.3	80.0	83.7	85.1	29.1	36.3	39.5	43.1	46.7	49.6
Coreset	13.4	18.8	24.3	29.4	34.6	31.1	58.4	76.1	83.7	86.0	87.7	29.1	36.2	41.1	44.4	47.4	51.2
Uncertainty	13.4	19.1	24.5	28.5	34.5	31.1	58.4	77.3	84.7	87.3	88.9	29.1	33.1	38.7	44.8	48.6	51.0
LLoss	13.4	18.3	23.5	29.6	34.7	31.1	53.6	74.4	81.6	86.1	88.2	29.1	35.9	40.0	44.3	49.2	51.1
SCAL	13.4	18.5	24.1	29.4	34.8	31.1	59.6	76.0	82.7	86.5	88.5	29.1	36.1	40.2	45.1	47.9	50.2
BADGE	13.4	19.0	24.2	28.0	31.3	31.1	61.0	75.4	83.2	86.8	88.3	29.1	35.5	41.8	47.3	51.1	54.2
BALD	13.4	19.7	24.6	29.4	34.8	31.1	60.3	76.2	82.8	85.9	87.6	29.1	36.6	41.5	45.5	48.3	50.4
CEC	13.4	19.6	25.2	29.1	33.0	31.1	61.3	76.9	83.7	86.3	88.7	29.1	36.0	41.0	45.8	49.0	51.3
HSSAL	13.4	19.4	24.5	29.1	33.5	31.1	58.5	77.4	84.8	87.3	88.9	29.1	**37.6**	42.3	45.8	49.5	52.3
TypiClust	13.4	20.4	25.3	30.7	35.8	31.1	65.2	74.9	79.1	83.3	86.6	29.1	36.8	42.0	46.5	49.8	52.6
Ours	**13.4**	**20.9**	**27.1**	**31.2**	**37.1**	**31.1**	**65.4**	**80.0**	**86.1**	**87.6**	**89.2**	**29.1**	37.4	**46.2**	**51.0**	**54.3**	**58.2**

post-initial budget, with final-budget gains of 1.3, 2.6, and 5.6 percentage points on CIFAR-100, SVHN, and Caltech-101. Comparing retrospectively with the best baseline at each budget gives the same ordering on CIFAR-100 and SVHN. On Caltech-101, HSSAL is 0.2 points higher at 1k labels, while Ours leads at later budgets. Per-cycle acquisition takes 33.54s, 28.29s, and 31.65s, respectively, with backbone feature extraction accounting for most of the time (Appendix B.3).

Dataset-dependent gains. SVHN shows stronger early gains and faster contraction of dominant residual directions, while CIFAR-100 and Caltech-101 retain persistent residual structure across later rounds. The residual dynamics and category-wise analyses below examine this relationship between acquisition gains and feature coverage. The long-tailed TinyImageNet evaluation further examines class coverage under an imbalanced training pool (Table 5).

We next analyze whether the performance gains are accompanied by meaningful changes in the operational nullspace residual space. We measure the global residual gap

$$E_{\mathrm{gap}}^{(t)} = \mathbb{E}_{x\sim\mathcal{D}_U^{(t)}}\left[\left\|P_{\mathrm{null}}^{(t)}h(x)\right\|_2^2\right], \tag{19}$$

which quantifies the average residual energy of unlabeled samples with respect to the current labeled-data subspace. We also compute the effective rank $R_{\mathrm{eff}}^{(t)}$ of the labeled-feature spectrum as a proxy for the spread of labeled-feature support. Full residual-gap and effective-rank curves are provided in Appendix B.2, and the corresponding gap-reduction rates and category-wise residual statistics are reported in Appendix B.4 and Appendix B.5.

Residual reduction is anisotropic. Figure 3 visualizes the evolution of E_{gap} and R_{eff}. On CIFAR-100, our method reduces the global residual gap from 0.65 to 0.24 and increases the effective rank from 102.5 to 338.2. On Caltech-101, it further reduces the residual gap from 0.52 to 0.12 while increasing the effective rank from 78.6 to 325.6. These trends suggest that selected samples broaden the spectral coverage of the labeled representation while reducing the average residual energy in the unlabeled pool. The more interesting finding is that this reduction is not uniform: simple and medium categories exhibit lower final gaps, while difficult categories retain larger residual values. Thus, active learning does not close the residual space uniformly; dominant residual directions are reduced earlier, while long-tail residual directions persist.

Figure 4 and Figure 5 provide complementary mechanism analyses. The category-level results show that residual reduction is anisotropic: although the global residual gap decreases, fixed representative difficult groups, selected before residual analysis based on non-ours baseline separability, still retain larger residual gaps, revealing a long-tail residual structure. The absorption-efficiency results further show that the same acquired samples are absorbed differently by different representations. ViT and CLIP achieve higher absorption efficiency and smaller forgetting on previously labeled samples, indicating that sample value depends on the current model state. Relative innovation and residual-space coverage track these changes during acquisition. Complete numerical results and the definition of representative category groups are provided in Appendix B.8, Appendix B.9, and Appendix B.1.

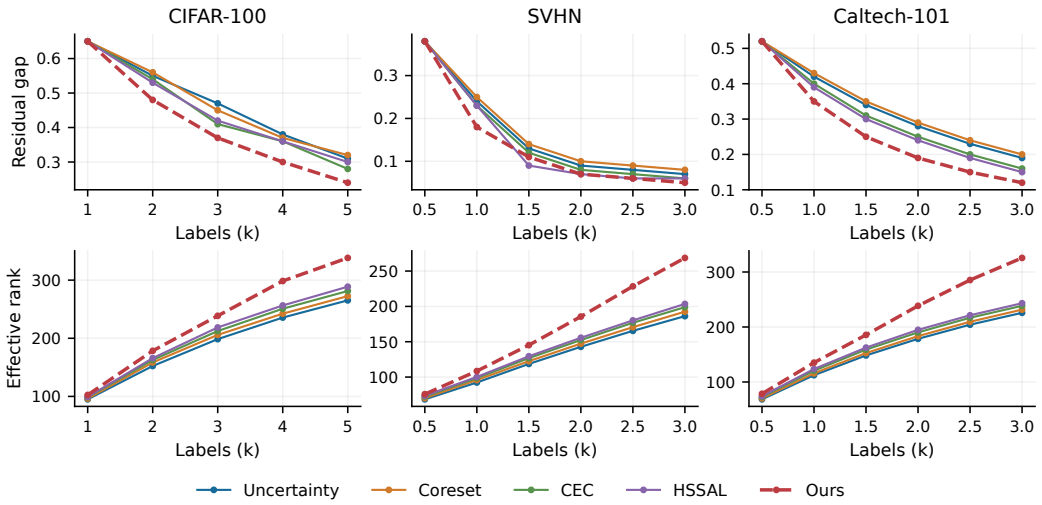


Figure 3: Evolution of the operational residual space. Top: average residual gap; bottom: effective rank of the labeled features. Each column shows a dataset-specific acquisition trajectory.

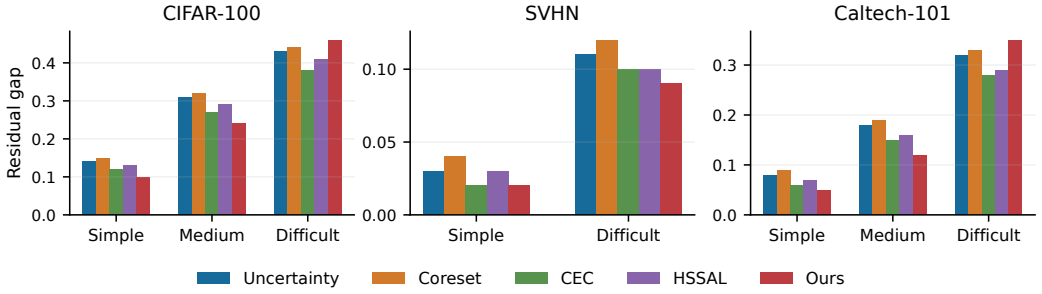


Figure 4: Category-wise residual anisotropy. The global residual gap decreases while difficult categories still retain persistent residual directions, revealing the long-tail structure of the residual space.

We further test whether sample value is intrinsic to the data or depends on the model state. Given the same unlabeled pool, we compute acquisition rankings under different backbones and measure their Spearman correlation and top-k overlap.

Sample value changes with the representation state. Table 2 shows that acquisition rankings have only low-to-moderate agreement across backbones, additional Spearman correlations, top-k overlaps, rank-difference statistics, and category-level analyses are provided in Appendix B.10. On CIFAR-100, ResNet-18 and CLIP obtain a Spearman correlation of only 0.34 and a Top-100 overlap of 31%. This result is more informative than a pure accuracy comparison: it shows that the same sample can have different acquisition value under different representation states. The effect is weaker on SVHN, where visual structure is simpler and model rankings are more aligned, whereas Caltech-101 shows stronger model-state dependence due to its heterogeneous category structure. These findings support the model-aware view that sample informativeness is not purely data-intrinsic, but depends on the current labeled representation and its operational nullspace residual space.

Ablation Study. We ablate the spectral-gap boundary, relative innovation, residual density, entropy, predicted loss, and greedy batch diversity. Table 3 summarizes the geometric ablations, and Table 4 reports the auxiliary-component trajectories with standard deviations. Full geometric trajectories are in Appendix B.6. For entropy and predicted-loss ablations, the removed weight is set to zero and the remaining weights are renormalized; the diversity ablation removes the distance bonus. Excluding the common initial point, removing entropy reduces average accuracy by 0.93/4.22/2.56 points on CIFAR-100/SVHN/Caltech-101. Removing predicted loss reduces it by 0.63/3.34/2.50 points, and

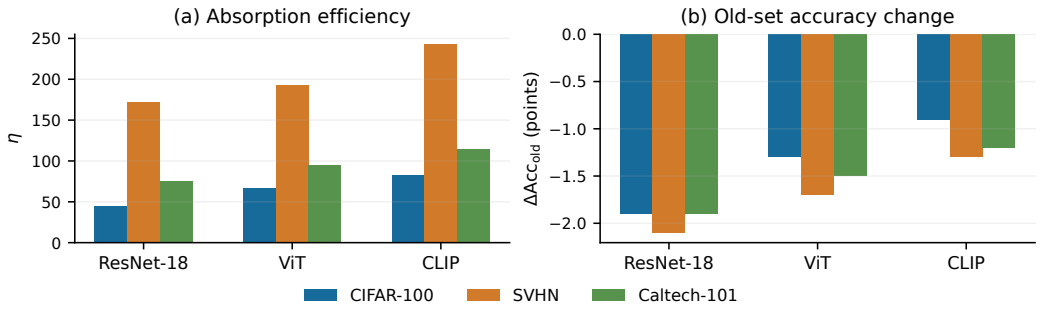


Figure 5: Model absorption capacity and stability. Left: absorption efficiency $\eta = \Delta\mathrm{Acc}/\Delta E_{\mathrm{gap}}$ across models and datasets. Right: forgetting measured by $\Delta\mathrm{Acc}_{\mathrm{old}}$ on previously labeled samples.

Table 2: Cross-model agreement of acquisition rankings under our method. Lower agreement indicates stronger model-state dependence of sample value.

Dataset	Model pair	Spearman ρ	Top-100	Top-500	Mean diff.
CIFAR-100	ResNet-18 vs ViT	0.41	38%	48%	125.3
CIFAR-100	ResNet-18 vs CLIP	0.34	31%	44%	158.7
CIFAR-100	ViT vs CLIP	0.49	45%	55%	98.4
SVHN	ResNet-18 vs ViT	0.53	52%	62%	68.4
SVHN	ResNet-18 vs CLIP	0.46	44%	56%	89.2
SVHN	ViT vs CLIP	0.60	58%	68%	52.6
Caltech-101	ResNet-18 vs ViT	0.39	35%	46%	142.6
Caltech-101	ResNet-18 vs CLIP	0.32	28%	41%	178.4
Caltech-101	ViT vs CLIP	0.46	42%	53%	112.3

removing diversity by 2.35/2.80/3.86 points. The auxiliary cues therefore contribute across the acquisition trajectory, with their relative importance varying by dataset. The geometric ablations show two small late-budget reversals: removing residual density raises CIFAR-100@5k by 0.1 point, and removing relative innovation raises Caltech-101@3k by 0.1 point. Residual density favors sparse regions, while relative innovation tracks persistence; either cue can become less informative as dominant residual modes are covered. The full method achieves the best average rank across the three datasets in the compact summary.

Table 3: Compact ablation summary. We report representative late-budget accuracy (%). Avg. rank summarizes the overall stability across datasets (lower is better).

Variant	CIFAR-100@5k	SVHN@3k	Caltech-101@3k	Avg. rank
Full method	<u>37.1</u>	**89.2**	<u>58.2</u>	**1.67**
w/o spectral-gap boundary	36.1	87.4	57.9	3.67
w/o relative innovation	36.6	<u>89.0</u>	**58.3**	2.00
w/o residual density	**37.2**	88.9	56.3	2.67

Long-tailed TinyImageNet. We evaluate class-imbalanced acquisition on TinyImageNet with an exponential long-tailed training pool (imbalance factor 50). All methods start from the same 2k labeled samples and acquire 2k samples per round until 10k labels. Many-, medium-, and few-shot groups correspond to the first, middle, and last thirds of the long-tailed class order; Appendix B.7 describes pool construction. Table 5 reports mean $\pm$ standard deviation over three seeds. Ours reaches $28.64 \pm 0.06\%$ overall accuracy and has the highest mean in each frequency group. Its few-shot accuracy is $9.79 \pm 2.45\%$, compared with $8.43 \pm 2.01\%$ for Coreset, with greater between-seed variation than in the overall accuracy.

Table 4: Component ablations on three benchmarks. Accuracy (%), mean ± standard deviation over three seeds. Each variant removes one component from the full method; the initial labeled point is shared.

CIFAR-100 Method	1k	2k	3k	4k	5k
Full	13.4 ± 0.61	20.9 ± 0.31	27.1 ± 0.73	31.2 ± 0.04	37.1 ± 0.39
w/o Entropy	13.4 ± 0.61	19.5 ± 0.22	26.2 ± 0.32	30.7 ± 0.20	36.2 ± 0.36
w/o Predicted Loss	13.4 ± 0.61	20.5 ± 0.45	26.6 ± 0.37	30.6 ± 0.56	36.1 ± 0.82
w/o Diversity	13.4 ± 0.61	18.1 ± 0.65	25.1 ± 1.03	28.9 ± 0.34	34.8 ± 1.02

SVHN Method	0.5k	1k	1.5k	2k	2.5k	3k
Full	31.1 ± 1.98	65.4 ± 0.37	80.0 ± 1.92	86.1 ± 1.35	87.6 ± 1.54	89.2 ± 0.81
w/o Entropy	31.1 ± 1.98	62.7 ± 1.71	73.2 ± 3.82	80.7 ± 2.51	84.5 ± 2.11	86.1 ± 1.74
w/o Predicted Loss	31.1 ± 1.98	63.8 ± 0.85	75.5 ± 1.56	81.1 ± 1.39	84.9 ± 1.93	86.3 ± 2.06
w/o Diversity	31.1 ± 1.98	61.8 ± 1.14	74.3 ± 1.09	83.0 ± 0.58	86.5 ± 0.74	88.7 ± 0.67

Caltech-101 Method	0.5k	1k	1.5k	2k	2.5k	3k
Full	29.1 ± 0.82	37.4 ± 1.41	46.2 ± 0.85	51.0 ± 0.59	54.3 ± 1.00	58.2 ± 0.93
w/o Entropy	29.1 ± 0.82	37.6 ± 1.20	41.5 ± 2.36	47.0 ± 1.72	51.4 ± 1.54	56.8 ± 1.07
w/o Predicted Loss	29.1 ± 0.82	37.7 ± 1.01	43.0 ± 1.55	48.0 ± 0.91	50.1 ± 1.92	55.8 ± 0.96
w/o Diversity	29.1 ± 0.82	36.1 ± 1.84	40.2 ± 2.09	45.8 ± 1.78	50.2 ± 1.49	55.5 ± 1.32

Table 5: Long-tailed TinyImageNet (imbalance factor 50), final 10k-label accuracy (%), mean ± standard deviation over three seeds. Many/Medium/Few denote the first/middle/last thirds of the long-tailed class order.

Method	Overall	Many	Medium	Few
TypiClust	27.46 ± 0.52	48.04 ± 1.25	26.12 ± 0.21	8.21 ± 3.00
BADGE	27.58 ± 0.74	47.81 ± 0.95	26.52 ± 0.13	8.39 ± 3.29
Coreset	27.38 ± 0.52	48.01 ± 0.06	25.65 ± 0.41	8.43 ± 2.01
Entropy	27.16 ± 0.30	47.51 ± 2.22	25.56 ± 1.56	8.37 ± 1.58
Ours	28.64 ± 0.06	48.60 ± 1.52	27.52 ± 0.77	9.79 ± 2.45

5 Conclusion

We introduced a nullspace-guided hybrid acquisition criterion that combines spectral residual geometry with entropy, predicted loss, and residual-space batch diversity. The experiments show improved sample efficiency in the evaluated settings, and the component ablations quantify the contributions of the fused cues. Tracking the labeled-feature spectrum reveals anisotropic residual contraction, persistent long-tail directions, and model-dependent acquisition rankings.

Limitations and future work. The operational nullspace describes residual directions within the learned representation, so its correspondence with missing semantic knowledge depends on representation expressiveness. A flat spectrum or an unstable spectral boundary can weaken the decomposition. Our evaluation focuses on heterogeneous, coverage-driven budgets; very-small-budget typicality regimes and later uncertainty-dominated regimes remain to be explored. Exact full-pool density entails pairwise computation, which can become costly at larger scale. Future work will test boundary stability under perturbations and evaluate residual directions against known generative factors (e.g., shape, color, or pose) omitted from the initial labels, using principal angles, projection-matrix distances, and residual-energy alignment.

Acknowledgments and Disclosures

Funding. This work received no external funding or other third-party support.

Competing interests. The authors declare no competing interests, including financial relationships outside the submitted work that could be perceived as influencing it.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[2] Burr Settles. Active learning literature survey. 2009.

[3] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54 (9):1–40, 2021.

[4] Sudhanshu Mittal, Joshua Niemeijer, Oezguen Cicek, Maxim Tatarchenko, Jan Ehrhardt, Jörg P Schäfer, Heinz Handels, and Thomas Brox. Realistic evaluation of deep active learning for image classification and semantic segmentation. *International Journal of Computer Vision*, 133 (7):4294–4316, 2025.

[5] David D Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm sigir forum*, volume 29, pages 13–19. ACM New York, NY, USA, 1995.

[6] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR, 2017.

[7] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

[8] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

[9] Jiangyi Wang and Na Zhao. Uncertainty meets diversity: A comprehensive active learning framework for indoor 3d object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20329–20339, 2025.

[10] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR, 2019.

[11] Chris Mingard, Guillermo Valle-Pérez, Joar Skalse, and Ard A Louis. Is sgd a bayesian sampler? well, almost. *Journal of Machine Learning Research*, 22(79):1–64, 2021.

[12] Hartmut Maennel, Ibrahim M Alabdulmohsin, Ilya O Tolstikhin, Robert Baldock, Olivier Bousquet, Sylvain Gelly, and Daniel Keysers. What do neural networks learn when trained with random labels? *Advances in Neural Information Processing Systems*, 33:19693–19704, 2020.

[13] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[14] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[15] Fei-Fei Li, Marco Andreetto, and Marc'Aurelio Ranzato. Caltech-101 object category dataset. Technical report, California Institute of Technology, 2004.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[18] 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*, pages 8748–8763. PMLR, 2021.

[19] Sheng-Jun Huang, Yi Li, and Ying-Peng Tang. Active learning for multiple target models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(1):140–154, 2026. doi: 10.1109/TPAMI.2025.3602682.

[20] Guy Hacohen, Avihu Dekel, and Daphna Weinshall. Active learning on a budget: Opposite strategies suit high and low budgets. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8175–8195. PMLR, 2022. URL `https://proceedings.mlr.press/v162/hacohen22a.html`.

[21] Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *International conference on machine learning*, pages 1954–1963. PMLR, 2015.

[22] Dongyuan Li, Zhen Wang, Yankai Chen, Renhe Jiang, Weiping Ding, and Manabu Okumura. A survey on deep active learning: Recent advances and new frontiers. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):5879–5899, 2025.

[23] Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. Batch active learning at scale. *Advances in Neural Information Processing Systems*, 34:11933–11944, 2021.

[24] Jiaxi Wu, Jiaxin Chen, and Di Huang. Entropy-based active learning for object detection with progressive diversity constraint. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9397–9406, 2022.

[25] Suraj Kothawade, Saikat Ghosh, Sumit Shekhar, Yu Xiang, and Rishabh Iyer. Talisman: targeted active learning for object detection with rare classes and slices using submodular mutual information. In *European Conference on Computer Vision*, pages 1–16. Springer, 2022.

[26] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117 (40):24652–24663, 2020.

[27] Tolga Birdal, Aaron Lou, Leonidas J Guibas, and Umut Simsekli. Intrinsic dimension, persistent homology and generalization in neural networks. *Advances in neural information processing systems*, 34:6776–6789, 2021.

[28] Ruili Feng, Kecheng Zheng, Yukun Huang, Deli Zhao, Michael Jordan, and Zheng-Jun Zha. Rank diminishing in deep neural networks. *Advances in Neural Information Processing Systems*, 35:33054–33065, 2022.

[29] Jialin Mao, Itay Griniasty, Han Kheng Teoh, Rahul Ramesh, Rubing Yang, Mark K Transtrum, James P Sethna, and Pratik Chaudhari. The training process of many deep networks explores the same low-dimensional manifold. *Proceedings of the National Academy of Sciences*, 121 (12):e2310002121, 2024.

[30] Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34:29820–29834, 2021.

[31] Can Yaras, Peng Wang, Zhihui Zhu, Laura Balzano, and Qing Qu. Neural collapse with normalized features: A geometric analysis over the riemannian manifold. *Advances in neural information processing systems*, 35:11547–11560, 2022.

[32] Lile Cai, Xun Xu, Lining Zhang, and Chuan-Sheng Foo. Exploring spatial diversity for region-based active learning. *IEEE Transactions on Image Processing*, 30:8702–8712, 2021.

[33] Zayd Hammoudeh and Daniel Lowd. Training data influence analysis and estimation: A survey. *Machine Learning*, 113(5):2351–2403, 2024.

[34] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

[35] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

[36] Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, 2020.

[37] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34:20596–20607, 2021.

[38] Zhaoxuan Wu, Yao Shu, and Bryan Kian Hsiang Low. Davinz: Data valuation using deep neural networks at initialization. In *International Conference on Machine Learning*, pages 24150–24176. PMLR, 2022.

[39] Haonan Wang, Wei Huang, Ziwei Wu, Hanghang Tong, Andrew J Margenot, and Jingrui He. Deep active learning by leveraging training dynamics. *Advances in Neural Information Processing Systems*, 35:25171–25184, 2022.

[40] Javad Zolfaghari Bengar, Bogdan Raducanu, and Joost van de Weijer. When deep learners change their mind: Learning dynamics for active learning. In *International Conference on Computer Analysis of Images and Patterns*, pages 403–413. Springer, 2021.

[41] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[42] Seungjae Hwang, Jaehoon Choi, and Junmo Choi. Uncertainty-based selective clustering for active learning. *IEEE Access*, 10:110983–110991, 2022.

[43] Bardia Safaei and Vishal M Patel. Active learning for vision-language models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4902–4912. IEEE, 2025.

[44] Wei Huang, Zhitong Xiong, Chenying Liu, and Xiao Xiang Zhu. Hierarchical semi-supervised active learning for remote sensing. *arXiv preprint arXiv:2511.18058*, 2025.

A Appendix: Theoretical and Implementation Details

This appendix develops the geometric properties of the spectral decomposition and provides implementation details for residual persistence and coverage-aware acquisition.

A.1 Subspace Decomposition and Operational Nullspace

Let $\widetilde{H}_L^{(t)} \in \mathbb{R}^{n \times d}$ denote the centered feature matrix of the labeled set at active learning cycle t, and let

$$\widetilde{H}_L^{(t)} = U_t \Sigma_t V_t^\top \tag{20}$$

be its singular value decomposition. Given an effective dimension k_t, we define the modeled subspace as

$$\mathcal{S}^{(t)} = \mathrm{span}(V_{t,1:k_t}), \tag{21}$$

where $V_{t,1:k_t}$ contains the leading right singular vectors. Its orthogonal complement is defined as

$$\mathcal{N}^{(t)} = \left(\mathcal{S}^{(t)}\right)^{\perp}. \tag{22}$$

In the proposed framework, $\mathcal{S}^{(t)}$ is treated as the set of feature directions most strongly supported by the current labeled data, while $\mathcal{N}^{(t)}$ is used as an operational nullspace proxy for residual directions that are weakly represented under this decomposition.

Lemma 1 (Orthogonal Decomposition). The feature space admits the direct-sum decomposition

$$\mathbb{R}^d = \mathcal{S}^{(t)} \oplus \mathcal{N}^{(t)}. \tag{23}$$

For any feature vector $h(x) \in \mathbb{R}^d$, there exists a unique decomposition

$$h(x) = h_{\parallel}^{(t)}(x) + h_{\perp}^{(t)}(x), \qquad h_{\parallel}^{(t)}(x) \in \mathcal{S}^{(t)}, \quad h_{\perp}^{(t)}(x) \in \mathcal{N}^{(t)}. \tag{24}$$

Proof. The selected right singular vectors are orthonormal. Their span and its orthogonal complement therefore yield two orthogonal subspaces whose direct sum is $\mathbb{R}^d$; an economy-size SVD need not return a full basis of the feature space. Existence and uniqueness follow from standard orthogonal decomposition. □

This result justifies decomposing each feature into a component represented by the current labeled subspace and a residual component in the corresponding orthogonal complement.

A.2 Projection Optimality and Residual Representation

The nullspace projection used in the method is

$$P_{\text{null}}^{(t)} = I - V_{t,1:k_t} V_{t,1:k_t}^{\top}, \tag{25}$$

and the corresponding residual representation is

$$z_{\perp}^{(t)}(x) = P_{\text{null}}^{(t)} h(x). \tag{26}$$

Theorem 1 (Projection Optimality). For any feature vector $h(x)$, the orthogonal projection onto $\mathcal{S}^{(t)}$ solves

$$P_{\mathcal{S}^{(t)}} h(x) = \arg\min_{y \in \mathcal{S}^{(t)}} \|h(x) - y\|_2^2. \tag{27}$$

The residual

$$P_{\mathcal{N}^{(t)}} h(x) = h(x) - P_{\mathcal{S}^{(t)}} h(x) \tag{28}$$

is therefore the minimum orthogonal residual of $h(x)$ with respect to the chosen modeled subspace.

Proof. Since $\mathcal{S}^{(t)}$ is a linear subspace with orthonormal basis $V_{t,1:k_t}$, the Euclidean projection onto this subspace is the unique minimizer of the squared reconstruction error. The residual is orthogonal to $\mathcal{S}^{(t)}$ by construction. □

The theorem identifies $z_{\perp}^{(t)}(x)$ as the reconstruction residual under the selected modeled subspace.

Corollary 1 (Residual under the Current Decomposition). The residual $z_{\perp}^{(t)}(x)$ is the orthogonal component of $h(x)$ relative to the modeled subspace $\mathcal{S}^{(t)}$. Its magnitude quantifies the degree to which $h(x)$ projects outside the current labeled-data subspace.

A.3 Residual Energy Decomposition

The orthogonal decomposition yields a natural energy decomposition.

Theorem 2 (Residual Energy Decomposition). For any feature $h(x)$,

$$\|h(x)\|_2^2 = \|P_{\mathcal{S}^{(t)}} h(x)\|_2^2 + \|P_{\mathcal{N}^{(t)}} h(x)\|_2^2. \tag{29}$$

Proof. The two projected components are orthogonal:

$$\langle P_{\mathcal{S}^{(t)}} h(x), P_{\mathcal{N}^{(t)}} h(x) \rangle = 0. \tag{30}$$

The result follows from the Pythagorean theorem. □

The term

$$\|P_{\mathcal{N}^{(t)}} h(x)\|_2^2 \tag{31}$$

is the residual energy of the sample with respect to the current labeled-data subspace. A large residual energy indicates that the sample has substantial feature content outside the current modeled subspace. This energy measures feature content outside the current modeled subspace.

A.4 Spectral Structure and Adaptive Boundary

The proposed method is spectrally grounded: both the modeled subspace and its operational nullspace are induced by the singular structure of labeled features.

Proposition 1 (Dominant Singular Directions). The subspace $\mathcal{S}^{(t)}$ spanned by the leading right singular vectors captures the principal variance directions of the centered labeled feature matrix $\widetilde{H}_L^{(t)}$.

Proof. By the variational characterization of singular value decomposition, the right singular vectors associated with larger singular values correspond to directions of larger projected variance in the labeled feature matrix. □

Proposition 2 (Adaptive Boundary via Spectral Gap). Let the normalized singular values be

$$\bar{\sigma}_i = \frac{\sigma_i}{\max_j \sigma_j + \epsilon}. \tag{32}$$

The rule

$$k_t = \arg \max_{i \in \{1,\dots,r-1\}} (\bar{\sigma}_i - \bar{\sigma}_{i+1}) \tag{33}$$

selects the largest spectral drop as an estimated transition point between dominant labeled directions and residual orthogonal directions.

The largest spectral drop determines the modeled boundary from the current labeled features, adapting the decomposition as the spectrum evolves across acquisition cycles.

A.5 Directional Nature of the Residual Representation

The residual representation is directional as well as magnitude-based. For a sample with nonzero residual, its normalized residual direction is

$$\tilde{z}_\perp^{(t)}(x) = \frac{z_\perp^{(t)}(x)}{\|z_\perp^{(t)}(x)\|_2}. \tag{34}$$

The magnitude $\|z_\perp^{(t)}(x)\|_2$ measures how strongly the sample projects outside the current labeled subspace, while $\tilde{z}_\perp^{(t)}(x)$ indicates the direction of this residual component.

This distinction is important for active learning. Two samples may have similar uncertainty or similar residual magnitude, but if their residuals point toward already covered residual directions, selecting both may provide limited additional coverage. Conversely, a sample with a distinctive residual direction may help cover a different region of the operational nullspace. This motivates both density-based residual scoring and diversity-aware batch selection in the main method.

A.6 Dynamics of the Operational Nullspace Across Cycles

As active learning proceeds, the labeled set changes and the modeled subspace is recomputed. Therefore, the operational nullspace $\mathcal{N}^{(t)}$ is generally cycle-dependent.

Proposition 3 (Conditional Subspace Contraction). If the representation map is fixed and the labeled-data span expands monotonically, then the corresponding orthogonal complement contracts:

$$\mathcal{S}^{(t)} \subseteq \mathcal{S}^{(t+1)} \quad \Longrightarrow \quad \mathcal{N}^{(t+1)} \subseteq \mathcal{N}^{(t)}. \tag{35}$$

Proof. For linear subspaces, inclusion reverses under orthogonal complementation: if $\mathcal{S}^{(t)} \subseteq \mathcal{S}^{(t+1)}$, then $(\mathcal{S}^{(t+1)})^\perp \subseteq (\mathcal{S}^{(t)})^\perp$. □

In practice, deep representations may drift across active learning cycles because the model is updated after acquiring new labels. Therefore, exact set inclusion is not guaranteed. We use this proposition only to explain the geometric tendency that, when the representation is locally stable and the labeled span expands, residual directions may be reduced over time. This also explains why absolute residual magnitudes are not directly comparable across cycles.

A.7 Relative Innovation as a Residual-Change Signal

The continuity adjustment used in the main method compares residual magnitudes across two consecutive acquisition cycles. Since backbone training continues after each acquisition round, the feature map is cycle-dependent. We therefore write the feature embedding explicitly as

$$h_t(x) = f_{\theta_t}(x), \tag{36}$$

where θ_t denotes the model parameters after training at cycle t.

For a remaining unlabeled sample $x \in \mathcal{D}_U^{(t)}$, the current residual magnitude is computed as

$$a_t(x) = \left\| P_{\text{null}}^{(t)} h_t(x) \right\|_2, \tag{37}$$

where $P_{\text{null}}^{(t)}$ is induced by the labeled-feature spectrum at cycle t. The previous residual magnitude is computed before the model update and cached as

$$a_{t-1}(x) = \left\| P_{\text{null}}^{(t-1)} h_{t-1}(x) \right\|_2. \tag{38}$$

Each residual magnitude uses the projector and feature representation from the same acquisition cycle.

The relative innovation score is then defined as

$$r^{(t)}(x) = \frac{a_t(x)}{a_{t-1}(x) + \epsilon}. \tag{39}$$

This ratio measures relative change rather than absolute residual strength. When $a_{t-1}(x) \gg \epsilon$, a value above one indicates an increase and a value near one indicates approximately unchanged magnitude. The latter can occur for either small or large residuals. Feature-scale changes between cycles can also affect the ratio.

The use of cached residual magnitudes has two practical advantages. First, it avoids storing the full previous-cycle residual vector for every unlabeled sample; only a scalar residual magnitude is required for computing $r^{(t)}(x)$. Second, it avoids direct vector-level comparison across different feature spaces. Since $h_{t-1}(x)$ and $h_t(x)$ may belong to different representation states, the method compares their cycle-wise residual strengths rather than their residual directions.

Because the scale of feature embeddings and nullspace residuals may vary across cycles, we apply rank-based normalization to the residual-derived scores before combining them with the other acquisition terms. The small constant ϵ prevents numerical instability when the previous residual magnitude is close to zero. In implementation, the first acquisition cycle has no previous residual cache; we therefore set $r^{(t)}(x) = 1$ for that cycle and rely on the current nullspace residual coverage score.

A.8 Coverage of Residual Directions

Residual magnitude alone is insufficient because many samples may concentrate around similar residual directions. To encourage coverage of under-explored regions, we estimate density in the nullspace residual space:

$$\rho^{(t)}(x) = \sum_{x' \in \mathcal{D}_U^{(t)}} \exp\left(-\frac{\|z_\perp^{(t)}(x) - z_\perp^{(t)}(x')\|_2^2}{\sigma^2 + \epsilon} \right). \tag{40}$$

The normalized density and sparsity-aware coverage score are

$$p^{(t)}(x) = \frac{\rho^{(t)}(x)}{\sum_{x' \in \mathcal{D}_U^{(t)}} \rho^{(t)}(x') + \epsilon}, \qquad c^{(t)}(x) = -\log\big(p^{(t)}(x) + \epsilon\big). \tag{41}$$

Proposition 5 (Coverage Interpretation). The score $c^{(t)}(x)$ assigns larger values to samples located in lower-density regions of the residual space. It therefore favors samples that are estimated to cover under-explored residual regions.

Combining $c^{(t)}(x)$ with the relative residual-change score gives

$$s_{\text{um}}^{(t)}(x) = r^{(t)}(x) c^{(t)}(x), \tag{42}$$

which favors samples with larger relative residual change in sparse residual regions.

A.9 Diversity-aware Batch Selection

The batch selection step balances individual acquisition value and residual directional diversity. Let $\mathcal{Q}$ denote the set of already selected samples in the current batch. The normalized distance from a candidate x to the current batch is

$$d_{\mathcal{Q}}^{(t)}(x) = \frac{\min_{x' \in \mathcal{Q}} \|z_{\perp}^{(t)}(x) - z_{\perp}^{(t)}(x')\|_2}{\max\limits_{u \in \mathcal{D}_U^{(t)} \setminus \mathcal{Q}} \min\limits_{x' \in \mathcal{Q}} \|z_{\perp}^{(t)}(u) - z_{\perp}^{(t)}(x')\|_2 + \epsilon}. \tag{43}$$

Each subsequent batch element is selected by

$$x^{\star} = \arg \max_{x \in \mathcal{D}_U^{(t)} \setminus \mathcal{Q}} \left[s^{(t)}(x) + \lambda_t d_{\mathcal{Q}}^{(t)}(x) \right], \tag{44}$$

where $\lambda_t = \lambda_{\text{div}} = 0.7$ for every cycle.

Proposition 6 (Greedy Residual-space Coverage). The diversity term $d_{\mathcal{Q}}^{(t)}(x)$ encourages the selected batch to spread across the nullspace residual space. Therefore, the greedy selection rule can be interpreted as balancing individual acquisition value with coverage of distinct residual regions.

This interpretation is consistent with the main method: active learning is viewed as a cycle-dependent process in which the current labeled set induces a spectral decomposition, residual persistence identifies directions that remain weakly covered, and batch selection improves coverage of these residual directions.

A.10 Geometric Structure of Normalized Residuals

When residual direction rather than magnitude is of interest, residuals can be normalized as

$$\tilde{z}_{\perp}^{(t)}(x) = \frac{z_{\perp}^{(t)}(x)}{\|z_{\perp}^{(t)}(x)\|_2 + \epsilon}. \tag{45}$$

Lemma 2 (Residual Direction Geometry). The normalized residuals approximately lie on a unit hypersphere, and their Euclidean distances reflect angular discrepancies between residual directions.

This supports the use of distance-based density estimation and diversity-aware batch selection in the residual space. In practice, normalization can be used when the goal is to emphasize directional discrepancy, while unnormalized residuals retain both magnitude and directional information.

Broader impacts. Reducing annotation requirements can make supervised learning more accessible. Acquisition on imbalanced or sensitive data can also amplify coverage gaps or biases if the representation fails to expose relevant groups; class and subgroup audits can help identify these effects alongside aggregate accuracy.

B Additional Experimental Results

This section provides supplementary numerical results omitted from the main text due to space limitations, including full residual dynamics, category-wise residual gaps, full-budget ablation studies, and model-dependent residual update analysis.

B.1 Additional Experimental Details

Datasets and backbones. We evaluate the proposed method on SVHN, CIFAR-100, and Caltech-101. For the main active learning comparison, we use Wide ResNet-18 on SVHN, ResNet-18 on CIFAR-100, and ImageNet-pretrained ResNet-50 on Caltech-101. For model-dependent analysis, we additionally evaluate ResNet-18, ViT, and CLIP-based representations.

Training protocol. All methods follow the same active learning protocol. At each acquisition cycle, a fixed number of samples is selected from the unlabeled pool and added to the labeled set. The backbone and loss-prediction head are initialized once per trial. After each acquisition round, their parameters are retained and training continues on the expanded labeled set; SGD optimizers and learning-rate schedulers are reinitialized at the start of each cycle. We use SGD with momentum 0.9, weight decay 5×10^{-3}, and an initial learning rate of 0.001 with a multi-step decay schedule. Unless otherwise specified, all methods share the same initialization, labeled budget, and training schedule.

Baselines. We compare with Random, Coreset, Uncertainty, LLoss, SCAL, BADGE, BALD, CEC, HSSAL, and TypiClust [20]. These methods cover random selection, uncertainty-based acquisition, diversity-based selection, loss-prediction-based acquisition, and recent active learning criteria.

Score implementation details. The loss-prediction head globally pools four intermediate backbone feature maps, applies a 128-dimensional linear layer and ReLU to each, concatenates the four outputs, and predicts one scalar loss. Its pairwise ranking objective uses the labeled per-sample classification losses [41]. The residual-density kernel uses bandwidth $\sigma = 1$. Component ranks are obtained from ascending sorting positions without tied-rank averaging; a singleton score maps to 1. These settings define the scores used in acquisition. The implementation uses $\epsilon = 10^{-8}$ for the residual ratio and caps ratios at their 95th percentile within the current unlabeled pool before computing the geometry score. The diversity coefficient is fixed at $\lambda_{\mathrm{div}} = 0.7$ across acquisition cycles.

Definition of Category Difficulty Groups For category-wise residual and model-state analyses, we use fixed representative class groups to examine whether residual reduction and model agreement vary across category difficulty. These groups are selected before computing the proposed residual statistics and are used only for analysis, not for sample acquisition or model training.

For CIFAR-100, we use automobile, ship, and chair as simple classes; cat, dog, and bird as medium classes; and mushroom, starfish, and snail as difficult classes. For SVHN, digits 0, 1, and 6 are treated as simple digits, digits 2, 3, 4, and 7 as medium digits, and digits 5 and 8 as difficult digits. For Caltech-101, Faces and Motorbikes are used as simple classes, Car_Side and Ketch as medium classes, and Flamingo and Brain as difficult classes.

The terms simple, medium, and difficult refer to fixed representative groups with different baseline separability and residual behavior, rather than to a learned grouping produced by the proposed method. This avoids circularity when analyzing category-wise residual anisotropy and model-state dependence.

B.2 Full Residual Dynamics

Tables 6–8 report the full residual gap and effective-rank curves on CIFAR-100, SVHN, and Caltech-101, respectively.

Table 6: Full residual dynamics on CIFAR-100. We report residual gap (lower is better) and effective rank (higher is better) across labeling budgets.

Metric	Method	1000	2000	3000	4000	5000
Residual gap						
	Uncertainty	0.65	0.55	0.47	0.38	0.31
	Coreset	0.65	0.56	0.45	0.37	0.32
	CEC	0.65	0.54	0.41	0.36	0.28
	HSSAL	0.65	0.53	0.42	0.36	0.30
	Ours	**0.65**	**0.48**	**0.37**	**0.30**	**0.24**
Effective rank						
	Uncertainty	95.0	152.3	198.5	235.6	265.2
	Coreset	96.2	158.4	205.8	242.3	272.5
	CEC	98.5	162.7	212.4	250.8	281.3
	HSSAL	99.2	165.8	218.6	256.4	288.7
	Ours	**102.5**	**178.4**	**238.6**	**298.5**	**338.2**

Table 7: Full residual dynamics on SVHN. We report residual gap (lower is better) and effective rank (higher is better) across labeling budgets.

Metric	Method	500	1000	1500	2000	2500	3000
Residual gap							
	Uncertainty	0.38	0.24	0.13	0.09	0.08	0.07
	Coreset	0.38	0.25	0.14	0.10	0.09	0.08
	CEC	0.38	0.23	0.12	0.08	0.07	0.06
	HSSAL	0.38	0.23	**0.09**	0.07	0.06	0.06
	Ours	**0.38**	**0.18**	0.11	**0.07**	**0.06**	**0.05**
Effective rank							
	Uncertainty	68.5	92.3	118.6	142.8	165.4	186.2
	Coreset	70.2	95.8	122.5	147.3	170.6	192.5
	CEC	72.4	98.6	126.8	152.4	176.8	198.9
	HSSAL	73.5	100.2	129.4	155.6	180.2	203.5
	Ours	**75.8**	**108.5**	**145.2**	**185.6**	**228.4**	**268.6**

Table 8: Full residual dynamics on Caltech-101. We report residual gap (lower is better) and effective rank (higher is better) across labeling budgets.

Metric	Method	500	1000	1500	2000	2500	3000
Residual gap							
	Uncertainty	0.52	0.42	0.34	0.28	0.23	0.19
	Coreset	0.52	0.43	0.35	0.29	0.24	0.20
	CEC	0.52	0.40	0.31	0.25	0.20	0.16
	HSSAL	0.52	0.39	0.30	0.24	0.19	0.15
	Ours	**0.52**	**0.35**	**0.25**	**0.19**	**0.15**	**0.12**
Effective rank							
	Uncertainty	68.5	112.4	148.3	178.5	204.2	225.8
	Coreset	70.2	115.8	152.6	183.4	209.5	231.6
	CEC	72.8	120.5	158.9	190.2	216.8	238.5
	HSSAL	74.2	123.6	162.5	194.8	221.3	243.2
	Ours	**78.6**	**135.2**	**185.6**	**238.5**	**285.4**	**325.6**

B.3 Runtime Analysis

Table 9 reports the per-cycle acquisition runtime of the proposed method. The reported time excludes model training and includes feature extraction, SVD-based subspace estimation, residual-density computation, and greedy batch selection. Feature extraction comprises backbone forward passes over the labeled and unlabeled pools to obtain embeddings and predictions, and accounts for most of the measured acquisition time. Totals also include score fusion and bookkeeping. Exact density computation evaluates all candidate pairs in GPU blocks.

B.4 Residual Gap Reduction Rates

Table 10 reports the gap-reduction and accuracy-gain statistics across the early and later acquisition stages.

B.5 Category-wise Residual Gap

Tables 11–13 provide category-wise residual gap statistics at the final budget. These results complement the main-text observation that residual reduction is anisotropic. For SVHN, Figure 4 and Table 12 show the simple and difficult groups as a two-group comparison; the medium group is included in the separate cross-backbone analysis in Appendix B.8.

The category-wise results are particularly informative: although the proposed method reduces the global residual gap most effectively, difficult categories can still retain higher residual gaps. This supports the main-text claim that active learning reduces the residual space anisotropically rather than uniformly.

Table 9: Per-cycle acquisition runtime of the proposed method. Runtime is measured with CUDA synchronization. The density term is evaluated as the exact full-pool Gaussian kernel sum, implemented in GPU blocks for memory efficiency rather than using subsampling or kNN approximation.

Dataset	Pool size	Feature extraction	SVD	Density	Greedy selection	Total
CIFAR-100	49,000	32.58 s	0.03 s	0.27 s	0.59 s	33.54 s
SVHN	72,757	26.72 s	0.03 s	0.83 s	0.63 s	28.29 s
Caltech-101	6,407	31.35 s	0.04 s	0.03 s	0.17 s	31.65 s

Table 10: Residual gap reduction rates and corresponding accuracy gains. For CIFAR-100 and SVHN, the early stage denotes rounds 0–2 and the later stage denotes rounds 2–4. For Caltech-101, the early stage denotes 500–1500 labels and the later stage denotes 1500–3000 labels.

Dataset	Method	Early ΔGap	Later ΔGap	Early ΔAcc	Later ΔAcc
CIFAR-100	Uncertainty	-0.18	-0.17	+11.8	+9.3
CIFAR-100	Coreset	-0.20	-0.19	+11.3	+9.0
CIFAR-100	CEC	-0.24	-0.18	+12.5	+9.6
CIFAR-100	HSSAL	-0.23	-0.17	+12.2	+9.5
CIFAR-100	**Ours**	**-0.28**	-0.18	**+13.7**	**+10.0**
SVHN	Uncertainty	-0.25	-0.07	+46.2	+11.6
SVHN	Coreset	-0.24	-0.04	+45.0	+11.6
SVHN	CEC	-0.26	-0.04	+45.8	+11.8
SVHN	HSSAL	**-0.29**	-0.03	+46.3	+9.7
SVHN	**Ours**	-0.27	**-0.06**	**+48.9**	+9.2
Caltech-101	Uncertainty	-0.18	-0.15	+9.6	+12.3
Caltech-101	Coreset	-0.17	-0.15	+12.0	+10.1
Caltech-101	CEC	-0.21	-0.15	+11.9	+10.3
Caltech-101	HSSAL	-0.22	-0.15	+13.2	+10.0
Caltech-101	**Ours**	**-0.27**	**-0.13**	**+17.1**	+12.0

Table 11: Category-wise residual gap on CIFAR-100 at the final budget.

Method	Simple classes	Medium classes	Difficult classes
Uncertainty	0.14	0.31	0.43
Coreset	0.15	0.32	0.44
CEC	0.12	0.27	0.38
HSSAL	0.13	0.29	0.41
Ours	**0.10**	**0.24**	0.46

Table 12: Category-wise residual gap on SVHN at the final budget.

Method	Simple digits	Difficult digits
Uncertainty	0.03	0.11
Coreset	0.04	0.12
CEC	**0.02**	0.10
HSSAL	0.03	0.10
Ours	**0.02**	**0.09**

Table 13: Category-wise residual gap on Caltech-101 at the final budget.

Method	Simple classes	Medium classes	Difficult classes
Uncertainty	0.08	0.18	0.32
Coreset	0.09	0.19	0.33
CEC	0.06	0.15	0.28
HSSAL	0.07	0.16	0.29
Ours	**0.05**	**0.12**	0.35

B.6 Full Ablation Results

Table 14 reports the full-budget ablation results. Each variant removes one component from the full method while keeping the rest of the protocol unchanged.

Table 14: Full ablation results on three benchmarks. Accuracy (%) is reported. For CIFAR-100, B1–B5 correspond to 1k, 2k, 3k, 4k, and 5k labels. For SVHN and Caltech-101, B1–B6 correspond to 0.5k, 1k, 1.5k, 2k, 2.5k, and 3k labels. "w/o SGB", "w/o RI", and "w/o RD" denote removing spectral-gap boundary, relative innovation, and residual density, respectively.

Variant	CIFAR-100					SVHN						Caltech-101					
	B1	B2	B3	B4	B5	B1	B2	B3	B4	B5	B6	B1	B2	B3	B4	B5	B6
Full	13.4	20.9	27.1	31.2	37.1	31.1	65.4	80.0	86.1	87.6	89.2	29.1	37.4	46.2	51.0	54.3	58.2
w/o SGB	13.4	20.0	26.6	30.2	36.1	31.1	63.9	79.7	84.4	87.6	87.4	29.1	37.4	46.1	50.2	54.0	57.9
w/o RI	13.4	20.3	27.1	30.5	36.6	31.1	65.3	80.0	86.1	87.5	89.0	29.1	37.2	46.2	50.6	53.8	58.3
w/o RD	13.4	20.5	27.0	30.8	37.2	31.1	64.0	79.0	85.3	87.4	88.9	29.1	37.1	44.7	49.7	54.1	56.3

These ablations show that the components interact with dataset structure. For example, residual density is most useful on Caltech-101, while individual components can occasionally match or slightly exceed the full method at a single budget. We therefore interpret ablations by their overall trends rather than by isolated budget points.

TypiClust comparison. All methods share the initial labeled set, candidate pool, and budget schedule within each dataset. The following table gives mean $\pm$ standard deviation over three seeds.

B.7 Long-tailed TinyImageNet Protocol

The 200 TinyImageNet classes are randomly ordered with seed 2022, and class rank $r \in \{0, \ldots, 199\}$ retains $n_r = \lfloor 500 \cdot 50^{-r/199} \rfloor$ training samples. This exponential profile yields 25,082 active-pool samples. The initial 2,000 labels include at least one sample per class when possible, with the remainder allocated using long-tailed weighting, leaving 23,082 initial unlabeled candidates. Each round acquires 2,000 samples until 10,000 labels. Many-, medium-, and few-shot groups are the first, middle, and last thirds of the class order, respectively.

B.8 Model-dependent Residual Update

We provide supplementary results on how different backbones update the residual space after acquiring the same batch of labeled samples. We report the initial accuracy, post-training accuracy, residual gap change, and descriptive residual-update efficiency $\eta = \Delta\mathrm{Acc}/(\Delta E_{\mathrm{gap}} + \epsilon)$.

B.9 Stability on Previously Labeled Samples

We also report the performance change on previously labeled samples after training with newly acquired labels. The results in Table 17 indicate that representation updates are architecture-dependent.

B.10 Additional Model-state Dependence Results

Tables 20 and 21 show that model agreement decreases from simple to difficult groups, while intra-group variance increases accordingly. This trend is more pronounced on CIFAR-100 and Caltech-101, suggesting that model-state dependence becomes stronger when the feature structure is more heterogeneous and the category-level residual geometry is less uniformly covered.

Table 15: TypiClust and Ours, accuracy (%), mean ± standard deviation over three seeds.

CIFAR-100					
Method	1k	2k	3k	4k	5k
TypiClust	13.4 ± 0.61	20.4 ± 0.26	25.3 ± 0.28	30.7 ± 0.47	35.8 ± 0.53
Ours	13.4 ± 0.61	20.9 ± 0.31	27.1 ± 0.73	31.2 ± 0.04	37.1 ± 0.39

SVHN						
Method	0.5k	1k	1.5k	2k	2.5k	3k
TypiClust	31.1 ± 1.98	65.2 ± 6.93	74.9 ± 1.79	79.1 ± 1.57	83.3 ± 1.87	86.6 ± 1.75
Ours	31.1 ± 1.98	65.4 ± 0.37	80.0 ± 1.92	86.1 ± 1.35	87.6 ± 1.54	89.2 ± 0.81

Caltech-101						
Method	0.5k	1k	1.5k	2k	2.5k	3k
TypiClust	29.1 ± 0.82	36.8 ± 1.05	42.0 ± 0.90	46.5 ± 1.05	49.8 ± 1.10	52.6 ± 1.41
Ours	29.1 ± 0.82	37.4 ± 1.41	46.2 ± 0.85	51.0 ± 0.59	54.3 ± 1.00	58.2 ± 0.93

Table 16: Model-dependent residual update under the proposed method.

Dataset	Model	Init. Acc	Final Acc	ΔAcc	Init. Gap	Final Gap	η
CIFAR-100	ResNet-18	13.4	20.9	+7.5	0.65	0.48	44.1
CIFAR-100	ViT	17.2	26.5	+9.3	0.59	0.45	66.4
CIFAR-100	CLIP	20.5	31.2	+10.7	0.54	0.41	82.3
SVHN	ResNet-18	31.1	65.4	+34.3	0.38	0.18	171.5
SVHN	ViT	36.2	71.0	+34.8	0.35	0.17	193.3
SVHN	CLIP	40.5	74.5	+34.0	0.32	0.18	242.9
Caltech-101	ResNet-18	29.1	41.8	+12.7	0.52	0.35	74.7
Caltech-101	ViT	32.5	47.6	+15.1	0.48	0.32	94.4
Caltech-101	CLIP	38.2	55.3	+17.1	0.44	0.29	114.0

Table 17: Stability analysis on previously labeled samples under the proposed method. We report old-set accuracy before and after training with newly acquired samples.

Dataset	Model	Before	After	$\Delta\text{Acc}_{\text{old}}$
CIFAR-100	ResNet-18	13.4	11.5	-1.9
CIFAR-100	ViT	17.2	15.9	-1.3
CIFAR-100	CLIP	20.5	19.6	-0.9
SVHN	ResNet-18	31.1	29.0	-2.1
SVHN	ViT	36.2	34.5	-1.7
SVHN	CLIP	40.5	39.2	-1.3
Caltech-101	ResNet-18	29.1	27.2	-1.9
Caltech-101	ViT	32.5	31.0	-1.5
Caltech-101	CLIP	38.2	37.0	-1.2

Table 18: Spearman rank correlations between acquisition rankings produced by different backbones. We report the correlations for multiple acquisition criteria when available.

Dataset	Model pair	Uncertainty	Coreset	CEC	HSSAL	Ours
CIFAR-100	ResNet-18 vs ViT	0.42	0.38	0.45	0.44	0.41
CIFAR-100	ResNet-18 vs CLIP	0.35	0.32	0.38	0.37	0.34
CIFAR-100	ViT vs CLIP	0.51	0.48	0.54	0.53	0.49
SVHN	ResNet-18 vs ViT	0.55	0.52	0.58	0.57	0.53
SVHN	ResNet-18 vs CLIP	0.48	0.45	0.51	0.50	0.46
SVHN	ViT vs CLIP	0.62	0.59	0.65	0.64	0.60
Caltech-101	ResNet-18 vs ViT	0.38	0.35	0.42	0.41	0.39
Caltech-101	ResNet-18 vs CLIP	0.31	0.28	0.35	0.34	0.32
Caltech-101	ViT vs CLIP	0.48	0.44	0.52	0.50	0.46

Table 19: Top-k overlap of acquisition rankings under the proposed method.

Dataset	Model pair	Top-100	Top-200	Top-500
CIFAR-100	ResNet-18 vs ViT	38%	42%	48%
CIFAR-100	ResNet-18 vs CLIP	31%	36%	44%
CIFAR-100	ViT vs CLIP	45%	49%	55%
SVHN	ResNet-18 vs ViT	52%	56%	62%
SVHN	ResNet-18 vs CLIP	44%	49%	56%
SVHN	ViT vs CLIP	58%	62%	68%
Caltech-101	ResNet-18 vs ViT	35%	39%	46%
Caltech-101	ResNet-18 vs CLIP	28%	33%	41%
Caltech-101	ViT vs CLIP	42%	46%	53%

Table 20: Distribution of acquisition-rank differences under the proposed method.

Dataset	Model pair	Mean	Std.	Max	Large-diff ratio
CIFAR-100	ResNet-18 vs ViT	125.3	98.2	487	18.5% (>250)
CIFAR-100	ResNet-18 vs CLIP	158.7	112.5	512	24.2% (>250)
CIFAR-100	ViT vs CLIP	98.4	85.6	423	12.8% (>250)
SVHN	ResNet-18 vs ViT	68.4	54.3	285	12.4% (>150)
SVHN	ResNet-18 vs CLIP	89.2	67.8	312	18.6% (>150)
SVHN	ViT vs CLIP	52.6	46.2	248	8.2% (>150)
Caltech-101	ResNet-18 vs ViT	142.6	105.4	512	21.5% (>250)
Caltech-101	ResNet-18 vs CLIP	178.4	128.6	548	28.4% (>250)
Caltech-101	ViT vs CLIP	112.3	92.8	468	15.2% (>250)

Table 21: Category-level model-state dependence under the proposed method. "R18–ViT" and "R18–CLIP" denote the Spearman rank correlation ρ between acquisition rankings produced by different backbones. "Var." denotes the intra-group variance.

Dataset	Group	R18–ViT	R18–CLIP	Var.
CIFAR-100	Simple	0.58	0.52	0.18
CIFAR-100	Medium	0.42	0.35	0.32
CIFAR-100	Difficult	0.28	0.21	0.45
SVHN	Simple	0.72	0.68	0.12
SVHN	Medium	0.58	0.52	0.24
SVHN	Difficult	0.38	0.31	0.38
Caltech-101	Simple	0.52	0.46	0.22
Caltech-101	Medium	0.38	0.31	0.35
Caltech-101	Difficult	0.24	0.18	0.48

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately describe the nullspace-guided hybrid active learning criterion, including entropy, predicted loss, and residual-space batch diversity and restrict the main claims to sample selection under limited annotation budgets. The claimed contributions are supported by the method design, theoretical analysis, ablation studies, and experiments on SVHN, CIFAR-100, Caltech-101, and a long-tailed TinyImageNet stress test.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of the proposed method, including its reliance on the quality and stability of learned representations, the computational cost of subspace estimation, and the empirical scope being limited to standard image classification active learning benchmarks. These limitations, together with the untested very-small and late-budget regimes, are discussed in the conclusion and appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions.
   - The authors should reflect on the scope of the claims made.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The theoretical results are presented with explicit assumptions, including representation stability and the evolution of the labeled-data subspace during active learning. Complete proofs and additional derivations are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material.

- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: The paper and appendix describe the main protocol, and the public repository supplies implementation and run instructions for CIFAR-100, SVHN, and Caltech-101. The cross-cycle training protocol and acquisition settings are specified, but the complete TinyImageNet-LT configuration and the definition used for the effective-rank statistics remain unspecified.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers.
   - If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
   - If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The implementation is publicly available at `https://github.com/chalesguo/nspaceal-release`. The repository includes source code, environment requirements, dataset preparation instructions, and commands for the main experiments on CIFAR-100, SVHN, and Caltech-101. These datasets are public benchmarks. Additional analyses and proofs are included in the appendix of this paper; no separate supplementary material is provided.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer.
   - The instructions should contain the exact command and environment needed to reproduce the results.
   - The authors should provide instructions on data access and preparation.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [No]

   Justification: The paper specifies the main backbones, labeling budgets, warm-start training protocol, optimizer, learning-rate schedule, score weights, fixed diversity coefficient, and residual-ratio clipping. The complete training and test configuration for TinyImageNet-LT is not reported.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The main accuracy table reports means over three seeds. The main-text component ablations and long-tailed TinyImageNet results report mean and standard deviation, and the appendix provides the same statistics for the Ours–TypiClust comparison. Other original baseline rows retain mean-only reporting. Standard deviations describe between-seed variability.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests.
   - The factors of variability that the error bars are capturing should be clearly stated.
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper reports acquisition runtimes, their component breakdown, and a synchronized timing protocol. GPU model and memory, complete training runtimes, and total compute are not reported.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research follows the NeurIPS Code of Ethics. The work uses public benchmark datasets and does not involve private data, sensitive personal information, human-subject data collection, or high-risk deployment.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses potential positive impacts, such as reducing annotation costs and improving data-efficient learning. It also discusses potential risks, including possible bias amplification if active learning is applied to sensitive or imbalanced real-world datasets without proper auditing.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models, generative models, scraped datasets, or systems with direct dual-use concerns. The proposed method is an active learning algorithm evaluated on public image classification benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards.
    - Datasets that have been scraped from the Internet could pose safety risks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the original sources of SVHN, CIFAR-100, and Caltech-101 and follows their public usage terms. No private, proprietary, or newly scraped datasets are used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license should be included for each asset when available.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released implementation is documented at `https://github.com/chalesguo/nspaceal-release`. Its README describes the code layout, environment setup, dataset preparation, and commands for the main experiments. The paper introduces no new dataset or pretrained model.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates.
    - At submission time, remember to anonymize your assets if applicable.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing, user studies, or new data collection from human subjects. All experiments are conducted on existing public benchmark datasets.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human-subject research, crowdsourcing, or collection of new personal data. Therefore, IRB approval or equivalent review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval or equivalent may be required for any human subjects research.
- For initial submissions, do not include any information that would break anonymity.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as an important, original, or non-standard component of the proposed method, experiments, or scientific claims. Any possible use of LLMs is limited to writing, editing, or formatting assistance and does not affect the core methodology or results.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to the NeurIPS LLM policy in the NeurIPS handbook for what should or should not be described.