Reliable Active Learning from Unreliable Labels via Neural Collapse Geometry

Atharv Goel*
IIIT Delhi
atharv21027@iiitd.ac.in

Sharat Agarwal*
IIIT Delhi
sharata@iiitd.ac.in

Saket Anand
IIIT Delhi
anands@iiitd.ac.in

Chetan Arora
IIT Delhi
chetan@cse.iitd.ac.in

Abstract

Active Learning (AL) promises to reduce annotation cost by prioritizing informative samples, yet its reliability is undermined when labels are noisy or when the data distribution shifts. In practice, annotators make mistakes, rare categories are ambiguous, and conventional AL heuristics (uncertainty, diversity) often amplify such errors by repeatedly selecting mislabeled or redundant samples. We propose Reliable Active Learning via Neural Collapse Geometry (NCAL-R), a framework that leverages the emergent geometric regularities of deep networks to counteract unreliable supervision. Our method introduces two complementary signals: (i) a **Class-Mean Alignment Perturbation score**, which quantifies how candidate samples structurally stabilize or distort inter-class geometry, and (ii) a **Feature Fluctuation score**, which captures temporal instability of representations across training checkpoints. By combining these signals, NCAL-R prioritizes samples that both preserve class separation and highlight ambiguous regions mitigating the effect of noisy or redundant labels. Experiments on ImageNet-100 and CIFAR100 show that NCAL-R consistently outperforms standard AL baselines, achieving higher accuracy with fewer labels, improved robustness under synthetic label noise, and stronger generalization to out-of-distribution data. These results suggest that incorporating geometric reliability criteria into acquisition decisions can make Active Learning less brittle to annotation errors and distribution shifts, a key step toward trustworthy deployment in real-world labeling pipelines.

1 Introduction

Deep learning depends on large-scale annotations (6), but real-world labels are often unreliable. This undermines Active Learning (AL) (9), whose heuristics (uncertainty (11; 4), diversity (9; 1)) can even exacerbate noise by selecting mislabeled, redundant, or ambiguous samples. This leads to inefficient label use, degraded generalization, and poor robustness under distribution shifts (3; 10).

Neural Collapse (NC) theory (7) shows that, late in training, features concentrate near class means, which align as a simplex ETF. These regularities provide stability even under imperfect supervision, suggesting that sample selection guided by NC dynamics could improve both efficiency and robustness (5; 2).

In this paper, we propose NCAL-R, a Neural Collapse–guided Active Learning framework designed for reliability under noisy or uncertain supervision. By quantifying how candidate samples perturb

*Equal contribution.

inter-class alignment and fluctuate across training checkpoints, NCAL-R selects points that both preserve feature structure and expose genuine ambiguities. Our experiments demonstrate improved accuracy with fewer labels, enhanced robustness to synthetic noise, and stronger out-of-distribution generalization.

2 Methodology

Problem Setting. We consider a standard pool-based Active Learning (AL) setting: a small labeled set $\mathcal{L}$, a large unlabeled set $\mathcal{U}$, and a model f_θ trained on $\mathcal{L}$. At each acquisition step, an AL strategy selects a batch $\mathcal{B} \subset \mathcal{U}$ for annotation. Our goal is to select $\mathcal{B}$ such that the learned representation is *robust* to both covariate shift and label drift, enabling improved in-distribution accuracy, OOD detection, and novel-class discovery.

Neural Collapse as a Structural Signal. In the late phase of training, deep classifiers often exhibit *Neural Collapse* (NC) (7): (NC1) within-class feature variance collapses, (NC2) class means form vertices of a simplex equiangular tight frame (ETF), (NC3) classifier weights align with class means, and (NC4) classification reduces to nearest-class-mean decisions. This emergent geometry reflects high class separability; deviations from it, or instability within it, may indicate *structurally valuable* samples that, when labeled, can improve generalization.

Acquisition Metrics. NCAL-R computes two complementary scores for each $x \in \mathcal{U}$:

1. **Class-Mean Alignment Perturbation (CMAP)**: Let μ_c denote the current empirical mean feature vector of class c and let $\hat{y}(x)$ be the model's predicted class for x. Denote by z the penultimate-layer feature for x. For any vector h we write $\bar{h} := h/\|h\|$ for its ℓ_2-normalized version. The class-mean updated by including z in class c (which has n_c members) is

$$\tilde{\mu}_c \;=\; \frac{n_c \mu_c + z}{n_c + 1}.$$

   Define the sum of normalized class means by $\bar{M} := \sum_{i=1}^{C} \bar{\mu}_i$, where C is the number of classes. We quantify the perturbation induced by x as the change in alignment of the (normalized) class mean with respect to the average of other class means:

$$\boxed{\mathrm{CMAP}(x) \;:=\; \big(\bar{\tilde{\mu}}_c - \bar{\mu}_c\big)^\top \big(\bar{M} - \bar{\mu}_c\big),} \tag{1}$$

   where $c = \hat{y}(x)$. Intuitively, δ_x measures how much adding x shifts its predicted-class mean toward (or away from) the centroid of the other class means; large positive values indicate samples that significantly perturb inter-class geometry and are therefore likely to refine decision boundaries. We derive this result in the Appendix.

2. **Feature Fluctuation (FF)**: Given model checkpoints $\{\theta_t\}_{t=T_i}^{T_f}$ where T_i and T_f are the start and end epochs in the NC phase, let $s_{\theta_t}(x) \in \mathbb{R}^c$ denote the pre-softmax logit vector produced for sample x. FF measures the variance of predicted logits for x across θ_t. High FF identifies samples with persistent uncertainty, even when most features have stabilized.

$$\boxed{\mathrm{FF}(x) \;=\; \sum_{t=T_i+1}^{T_f} \mathbf{1}\big[\arg\max s_{\theta_t}(x) \neq \arg\max s_{\theta_{t-1}}(x)\big]} \tag{2}$$

Combined Acquisition Strategy. NCAL-R selects the top-k samples from $\mathcal{U}$ by ranking CMAP and FF separately, standardizing each by their mean and standard deviation, and averaging:

$$\mathrm{Score}(x) = \frac{\mathrm{CMAP}(x) + \mathrm{FF}(x)}{2}.$$

This yields a batch $\mathcal{B}$ that contains both structurally impactful and prediction-unstable samples, shaping the representation to be both discriminative and adaptable. NCAL-R requires no auxiliary networks, pseudo-labeling, or task-specific tuning, and can be applied to any backbone or modality where feature embeddings can be extracted.

3 Experiments

Experimental Setup. We evaluate NCAL-R on tasks including classification, OOD detection, OOD generalization, and general category discovery. Label drift is tested under the GCD protocol; covariate shift via linear probes on OOD datasets. Unless noted, we use a ResNet-18 backbone, 5% acquisition per cycle, and compare to Random, CoreSet (9), and CDAL (1).

Evaluation Metrics. We report: (i) **All-class accuracy**: top-1 classification accuracy over both known and novel classes; (ii) **Novel-class accuracy**: GCD accuracy restricted to novel classes; (iii) **Known-class accuracy**: classification accuracy on known classes; (iv) **AUROC** for binary OOD detection between in-distribution and OOD samples.

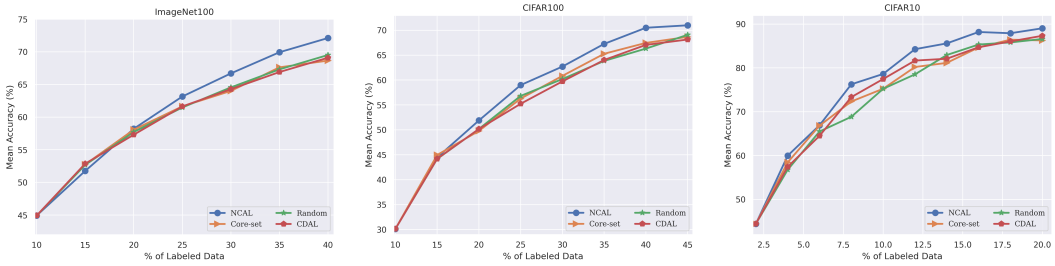


Figure 1: Comparison of test accuracy across varying label budgets on three benchmark datasets—ImageNet100, CIFAR100, and CIFAR10. NCAL's good performance even at lower annotation budgets suggests that its Neural Collapse-guided selection promotes more structured and representative feature learning. *(Note: accuracy for 100% data of ImageNet100, CIFAR100 and CIFAR10 are: 79.16%, 70.75% and 90% respectively. Reported results are average of 3 independent runs.)*

Method	10%	15%	20%	25%	30%	35%	40%	100%
Random	77.18	80.57	84.13	85.45	86.89	87.82	88.67	93.68
CDAL		81.78	84.28	85.9	86.34	87.98	88.92	
Coreset		81.56	83.73	85.66	87.1	88.29	88.95	
NCAL		**82.49**	**85.55**	**87.89**	**89.15**	**90.53**	**91.53**	

Table 1: AUROC scores for Far-OOD detection on the OpenImage-O dataset trained on ImageNet-100 with varying annotation budgets.

Method	All Classes	Old Classes	New Classes	Val Accuracy
Random	33.20	50.34	20.35	36.20
CDAL	33.39	49.96	20.96	36.94
Coreset	32.23	49.98	18.92	36.44
NCAL	**35.07**	**51.95**	**23.05**	**37.76**

Table 2: Performance across all, old, and new classes along with validation accuracy.

Covariate Shift Results. We test the ability to generalize to OOD datasets by training a linear probe over the learned embeddings. Table 4 shows that NCAL-R improves OOD classification by $\sim 2\%$ on average across 8 varying datasets, over all baselines. This demonstrates the adaptability of NCAL-R's feature space to both NearOOD and FarOOD scenarios.

Label Drift and GCD. NCAL-R's geometry-aware selection yields features that support unsupervised novel-class discovery while maintaining high accuracy on known classes. In the GCD setting with 60-40 split, NCAL-R improves novel-class accuracy by $+2.1$ points over the best baseline without supervision on novel classes, and by $+1.6$ points on known classes. This demonstrates that NCAL-R's feature space is inherently adaptable to evolving label spaces, without forgetting past label information.

Method	10%	15%	20%	25%	30%	35%	40%	100%
Random	77.18	80.57	84.13	85.45	86.89	87.82	88.67	93.68
CDAL		81.78	84.28	85.9	86.34	87.98	88.92	
Coreset		81.56	83.73	85.66	87.10	88.29	88.95	
NCAL		**82.49**	**85.55**	**87.89**	**89.15**	**90.53**	**91.53**	

Table 3: AUROC scores for Far-OOD detection on the OpenImage-O dataset trained on Imagenet-100 with varying annotation budgets.

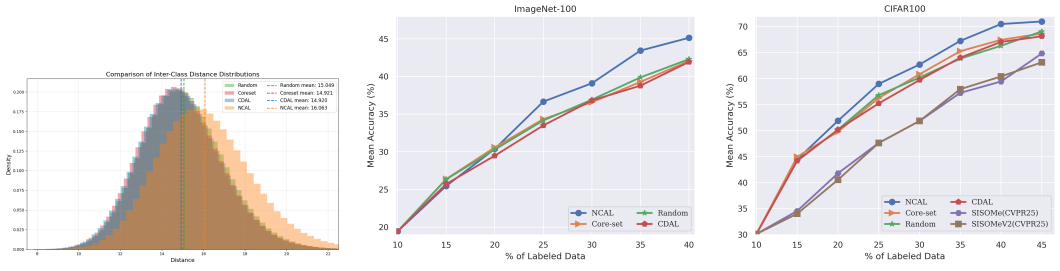
	Val / Train Acc		OOD Generalization (linear probe val accuracy)								
	Val (%)	Train	ImgNet-R	CIFAR100	Flowers	NINCO	CUB	Aircraft	Pets	STL	Avg
Random	69.51	96.44	18.06	41.64	58.69	64.23	37.84	15.26	42.34	68.67	46.95
CDAL	69.09	96.55	17.56	41.98	58.13	65.87	38.53	15.03	42.65	68.27	47.21
Coreset	68.65	96.42	16.93	42.02	57.96	65.11	37.86	15.15	42.22	68.68	47.00
NCAL	**72.11**	**95.22**	**19.27**	**43.78**	**60.87**	**67.66**	**40.01**	**15.38**	**44.70**	**70.49**	**48.98**
100%	79.16	95.27	20.01	45.31	61.77	69.90	42.29	19.08	46.14	71.45	50.87

Table 4: Comparison of validation accuracy, Neural Collapse metrics, and OOD generalization (measured via linear probe accuracy) across multiple benchmarks. `NCAL` consistently achieves stronger generalization to diverse OOD datasets compared to baselines.

Inter-Class Separation in Feature Space: To further analyze the structure of learned representations, we examine the distribution of inter-class distances in the penultimate feature space. fig. 2a shows a density plot comparing these distributions across different Active Learning strategies. Notably, `NCAL-R` exhibits a clear rightward shift, indicating larger average separation between class centroids (mean = 15.944), compared to Random (15.114), Coreset (15.070), and CDAL (15.130). This increased inter-class distance suggests that `NCAL-R` promotes more discriminative and geometrically separated class representations an essential property for improving generalization, especially under low-label regimes and OOD scenarios.

Performance Comparison in Long-Tail Distribution: Real-world data comes in a long-tail distributions, leading to bias towards certain classes. We construct a highly imbalanced version of ImageNet-100 by applying an exponential decay to class sample counts with a decay factor of $\beta = 0.05$, leading to a pool of 41,454 samples. An active-learning cycle with this pool achieves 45.15% for `NCAL-R`, compared to 42.30% (Random), 42.06% (Coreset) and 41.94% (CDAL) an improvement of +3% with only 16k images fig. 2b.

Evaluating Transferability of ActiveOOD Strategies: In this ablation, we evaluate the recently proposed ActiveOOD technique SISOMe (8) for Open-Set in our Closed-Set AL setup by removing its OOD filtering component. As shown in fig. 2c, SISOMe performs significantly worse than both standard baselines and NCAL. These results indicate that SISOMe's scoring heuristics do not transfer well to settings without explicit OOD filtering.



(a) Inter-Class Separation in Feature Space

(b) Comparison in Long-Tail Distribution

(c) Comparison with ActiveOOD

Figure 2: Ablation

4 Conclusion

We presented NCAL-R, an Active Learning framework that leverages Neural Collapse geometry. By combining CMAP and FF scores, NCAL selects structurally informative and uncertain samples, yielding more discriminative and robust feature spaces. Experiments show consistent gains across accuracy, OOD detection, OOD generalization, and category discovery. At its core, NCAL-R shows that structure matters – aligning acquisition decisions with the emergent geometry of deep networks can pay significant dividends.

References

[1] Agarwal, S., Arora, H., Anand, S., Arora, C.: Contextual diversity for active learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16. pp. 137–153. Springer (2020)

[2] Ammar, M.B., Belkhir, N., Popescu, S., Manzanera, A., Franchi, G.: Neco: Neural collapse based out-of-distribution detection. In: ICLR (2024), https://openreview.net/forum?id=9ROuKblmi7

[3] Chitta, K., Berman, M., Loy, C.C., Schiele, B.: Training dynamics for sample selection in deep learning. In: Advances in Neural Information Processing Systems. vol. 34, pp. 17967–17979 (2021)

[4] Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. In: Proceedings of the 34th International Conference on Machine Learning. pp. 1183–1192 (2017)

[5] Haas, J., Yolland, W., Rabus, B.T.: Linking neural collapse and l2 normalization with improved out-of-distribution detection in deep neural networks. Transactions on Machine Learning Research (2024)

[6] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)

[7] Papyan, V., Han, X., Donoho, D.: Prevalence of neural collapse during the terminal phase of deep learning training. Proceedings of the National Academy of Sciences **117**(40), 24652–24663 (2020)

[8] Schmidt, S., Schenk, L., Schwinn, L., Günnemann, S.: Joint out-of-distribution filtering and data discovery active learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25677–25687 (2025)

[9] Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. In: International Conference on Learning Representations (2018)

[10] Wang, Z., Hu, Z., Xu, Z., et al.: Caffe: Category-aware feature feedback for efficient active learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6578–6588 (2022)

[11] Yoo, D., Kweon, I.S.: Learning loss for active learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 93–102 (2019)

A Deriving the CMAP

In this section, we derive the Class-mean Alignment Perturbation (CMAP) score (δ_x), as introduced in Sec. 2. The CMAP quantifies the change in alignment of the normalized class means induced by candidate sample x.

Let x be a candidate sample with penultimate-layer feature embedding z, and let $c = f(x)$ be its predicted class. Denote C as the number of classes. Suppose the current class mean for class c is μ_c, computed over n_c training samples. If x is added to class c, the updated class mean becomes:

$$\tilde{\mu}_c = \frac{n_c \mu_c + z}{n_c + 1}$$

Let $\mu_1, \ldots, \mu_k$ be the class means before adding z, and $\bar{\mu}_i = \mu_i / \|\mu_i\|$ be their ℓ_2-normalized versions. Define the sum of all normalized class means: $\bar{M} := \sum_{i=1}^{C} \bar{\mu}_i$

Let CMA_{init} and CMA_{final} denote the Class-mean alignment (CMA) before and after adding the sample, respectively. This expression is the pair-wise average cosine similarity of class means. Then,

$$
\begin{aligned}
CMA_{init} &= \frac{1}{k(k-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{k} \text{Sim}(\mu_i, \mu_j) \\
&= \frac{1}{k(k-1)} \left[\sum_{\substack{i,j=1 \\ i \neq j \\ i \neq c \\ j \neq c}} \text{Sim}(\mu_i, \mu_j) + 2 \sum_{\substack{i=1 \\ i \neq c}}^{k} \text{Sim}(\mu_i, \mu_c) \right]
\end{aligned}
$$

We isolate the terms involving class c since only those are affected by the perturbation. The remaining terms cancel when computing the delta:

$$
\begin{aligned}
\delta_x &= CMA_{\text{final}} - CMA_{\text{init}} \\
&= \frac{2}{k(k-1)} \sum_{\substack{i=1 \\ i \neq c}}^{k} [\text{Sim}(\tilde{\mu}_c, \mu_i) - \text{Sim}(\mu_c, \mu_i)]
\end{aligned}
$$

Using cosine similarity, $\text{Sim}(a, b) = \frac{a^T b}{\|a\| \|b\|}$, and denoting $\bar{\mu} = \frac{\mu}{\|\mu\|}$ as the unit-norm version of a vector, we simplify the expression:

$$
\begin{aligned}
\delta_x &= \frac{2}{k(k-1)} \sum_{\substack{i=1 \\ i \neq c}}^{k} \left[\bar{\tilde{\mu}}_c^T \bar{\mu}_i - \bar{\mu}_c^T \bar{\mu}_i \right] \\
&= \frac{2}{k(k-1)} \sum_{\substack{i=1 \\ i \neq c}}^{k} \left[(\bar{\tilde{\mu}}_c - \bar{\mu}_c)^T \bar{\mu}_i \right] \\
&= \frac{2}{k(k-1)} (\bar{\tilde{\mu}}_c - \bar{\mu}_c)^T \sum_{\substack{i=1 \\ i \neq c}}^{k} \bar{\mu}_i \\
&= \frac{2}{k(k-1)} (\bar{\tilde{\mu}}_c - \bar{\mu}_c)^T (\bar{M} - \bar{\mu}_c)
\end{aligned}
$$

Finally, omitting the constant for interpretability and ranking purposes, we define the perturbation score:

$$
\boxed{\text{CMAP}(x) := \delta_x = (\bar{\tilde{\mu}}_c - \bar{\mu}_c)^T (\bar{M} - \bar{\mu}_c)}
$$

Implementation note: CMAP requires only the current per-class counts $\{n_c\}$ and means $\{\mu_c\}$ plus the feature z for x; the increment $\tilde{\mu}_c$ can be computed cheaply and $\bar{M}$ updated incrementally if desired.

B Training Protocol

At each AL cycle:

1. Train f_θ on $\mathcal{L}$ until the Neural Collapse phase.
2. Compute CMAP and FF for all $x \in \mathcal{U}$.
3. Select $\mathcal{B}$ using the combined score, query labels, and update $\mathcal{L} \leftarrow \mathcal{L} \cup \mathcal{B}$.
4. Repeat until budget is exhausted.

Experimental settings.

1. **ImageNet100:** Initial pool consists of 10% randomly sampled data, i.e. 13,000 samples. In each iteration, we select 5% (i.e., 6,500) samples to be annotated and added to the pool for next iteration of training. We terminate the loop when our labelled pool reaches 40% of the training set.
2. **CIFAR100:** The initial pool size is 10%, i.e. 5,000 images, acquiring 5% (2,500 images) in each cycle. We terminate at 45% pool size.
3. **CIFAR10:** The initial pool is 2% (i.e. 1,000 images), acquiring 2% images every cycle until 20% pool size.

Compute. We run all our experiments on an A100 GPU with a 20 GB memory capacity.

C Algorithm Pseudo Code

Input: Unlabeled pool $\mathcal{U}$, labeled set $\mathcal{L}$, class means $\{\mu_c\}$, model checkpoints $\{f_t\}_{t=T_i}^{T_f}$, acquisition budget k
Output: Selected sample indices $\mathcal{A} \subset \mathcal{U}, |\mathcal{A}| = k$
Initialize empty lists $\{\delta_x\}$ and $\{\phi_x\}$ for each $x \in \mathcal{U}$
Compute normalized class means $\bar{\mu}_c := \mu_c / \|\mu_c\|$ for each class c
Compute $\bar{M} := \sum_c \bar{\mu}_c$
for all $x \in \mathcal{U}$ **do**
  $c \leftarrow f(x)$ {Predicted label for x}
  $z \leftarrow$ penultimate-layer feature of x
  $\tilde{\mu}_c \leftarrow \frac{n_c \mu_c + z}{n_c + 1}$
  $\tilde{\bar{\mu}}_c \leftarrow \tilde{\mu}_c / \|\tilde{\mu}_c\|$
  $\delta_x \leftarrow (\tilde{\bar{\mu}}_c - \bar{\mu}_c)^T (\bar{M} - \bar{\mu}_c)$
  $\phi_x \leftarrow 0$
  for $t = T_i + 1$ **to** T_f **do**
    if $f_t(x) \neq f_{t-1}(x)$ **then**
      $\phi_x \leftarrow \phi_x + 1$
    end if
  end for
end for
Standardize scores using Z-score normalization:

$$\mathrm{CMAP}(x) \leftarrow \frac{\delta_x - \mu_\delta}{\sigma_\delta}, \quad \mathrm{FF}(x) \leftarrow \frac{\phi_x - \mu_\phi}{\sigma_\phi}$$

Compute acquisition scores: $s_x := \frac{\mathrm{CMAP}(x) + \mathrm{FF}(x)}{2}$ for each $x \in \mathcal{U}$
Select top-k samples: $\mathcal{A} \leftarrow \texttt{TopK}(\{s_x\}_{x \in \mathcal{U}}, k)$
return $\mathcal{A}$

Algorithm 1: NCAL Acquisition Function

D Limitations of NCAL-R

Limitations. NCAL-R relies on models being trained into the *neural collapse* regime, i.e., the terminal phase where training accuracy plateaus and geometric regularities emerge. Reaching this phase can require many epochs, depending on the dataset and architecture, which may limit efficiency. Moreover, the study of Neural Collapse in large-scale models (e.g., LLMs) remains limited. Since such models are typically trained for only a few epochs, it is unclear whether NCAL-R's assumptions hold in these settings. We have not evaluated NCAL-R under such large-scale regimes, and adapting it there may require further investigation.

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 claim that NCAL-R improves Active Learning robustness by (i) selecting samples that perturb inter-class geometry and (ii) identifying prediction-unstable samples. Our experiments on CIFAR-100 and ImageNet-100 directly evaluate these points, showing consistent gains in accuracy, robustness to synthetic label noise, and OOD generalization. We avoid over-claiming: we do not extend results to very large-scale models (see Limitations), so the scope stated matches the contributions delivered.

   Guidelines:

   - The answer NA 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. A No or NA answer to this question will not be perceived well by the reviewers.
   - 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: We include a dedicated Limitations section in the Appendix. In particular, we note that NCAL-R requires models to be trained to neural collapse, which can be computationally expensive for some architectures and datasets. We also acknowledge that we have not tested our method on very large-scale models such as LLMs, where neural collapse behavior is less studied and may not emerge under typical training schedules. These points transparently delimit the scope of our contributions.

   Guidelines:

   - The answer NA 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 (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

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: While we do not propose any new theorems or lemmas, the CMAP computes a score based on results derived from theorems in related work. We have derived the result from first principles in the Appendix.

   Guidelines:

   - The answer NA 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, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or 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: [Yes]

   Justification: Our main contribution is an algorithm, and all information required to implement and reproduce it are provided. To reproduce our exact experiments and results, we release all our code with scripts for each experiment. We make it easy to reproduce and use our results. We also provide the pseudocode for our algorithm in the Appendix.

   Guidelines:

   - The answer NA 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: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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: All our experiments are fully reproducible, and instructions to do so are documented in the GitHub readme with exact commands to be run to invoke each script. The full, heavily documented codebase will be released upon acceptance of this paper.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

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, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All experiment setting details relevant to our algorithm are described in the paper. More fine-grained details are in the Appendix. Details on exact hyperparameters, splits, etc. are documented in the code repository.

   Guidelines:

   - The answer NA 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 and make sense of them.
   - 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: Active Learning performance can be sensitive to random initialization and acquisition order. To mitigate this variance, we repeat all experiments with 3 different random seeds and report the mean performance across runs. Averaging across seeds provides a fairer estimate of the underlying performance and ensures the reported trends are statistically reliable.

   Guidelines:

   - The answer NA 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, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

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: [Yes]

   Justification: We include the compute requirements and implementation details in the Appendix.

   Guidelines:

   - The answer NA 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.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

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: We have thoroughly reviewed the code of ethics and evaluated our work against it; we are fully compliant.

Guidelines:

- The answer NA 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 (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [NA]

    Justification: Our work poses no potential risks or impacts to society. We see no direct application to any malicious use cases. Our work improves the generalization of ML systems under unreliable data, it does not lead to direct application or deployment.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work has no such risks.

    Guidelines:

    - The answer NA 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 to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

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: [NA]

    Justification: We do not use any existing assets.

    Guidelines:

    - The answer NA 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 (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [NA]

    Justification: We don't introduce any new assets. All code needed to reproduce our experiments is in the GitHub repository with proper documentation.

    Guidelines:

    - The answer NA 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. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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: [NA]

    Justification: This work does not involve any human subjects or crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - 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: [NA]

    Justification: This work does not involve any human subjects or crowdsourcing.

    Guidelines:

    - The answer NA 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. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

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 rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Our work does not use LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.