Conformal Cross-Modal Active Learning

Supplementary Material

A. Implementation details

CIFAR100 contains 60K natural images (32×32) across 100 fine-grained categories, providing a standard benchmark for image classification under limited-label regimes. **Food101** consists of 101K high-resolution food images with substantial intra-class variation. **DomainNet-Real** is a subset of the large-scale DomainNet benchmark, with real-world images from 345 categories, and is used to assess robustness and cross-domain generalization. **Caltech101** includes ∼9K images from 101 object categories with relatively clean backgrounds, while **Caltech256** extends this to 30K images over 256 categories with greater intra-class diversity. At each active learning round, the acquisition batch size is set to the number of classes: $B = 100$ for CIFAR100, $B = 101$ for Food101, and $B = 345$ for DomainNet-Real. We follow a fully *cold-start* protocol [2]: the student classifier is reinitialized and retrained from scratch at every acquisition step. Both the VLM teacher and the student's vision backbone remain frozen throughout; features for all datasets are extracted once via a single forward pass and subsequently cached for all experiments.

B. Text feature extraction from VLMs

For all experiments, the teacher model is a frozen CLIP ViT-L/14 [33] encoder that provides both image and text representations. To obtain class-wise text embeddings, we follow the CLIP text encoder pipeline and construct prompts for each dataset category using multiple natural language templates. Given a class name c, we build prompts of the form "a photo of a {class}", "a bright photo of a {class}", etc. Each prompt is tokenized and encoded with the CLIP text encoder, and the final class embedding is obtained by averaging the normalized embeddings across all templates.

Prompt templates. Unless otherwise stated, we use the following default set:

```
‘‘a photo of a {class}’’
‘‘a photo of the {class}’’
‘‘a blurry photo of a {class}’’
‘‘a photo of one {class}’’
‘‘a close-up photo of a {class}’’
‘‘a rendition of a {class}’’
‘‘a bright photo of a {class}’’
‘‘a low resolution photo of a {class}’’
‘‘a cropped photo of a {class}’’
‘‘a clean photo of a {class}’’
```

These templates follow the standard CLIP zero-shot classification protocol [33], ensuring semantically rich and diverse text representations.

Dataset-specific class mappings. To handle datasets with different class label formats (e.g., numeric labels or textual names), we extract class names using a unified loader:

- **CIFAR100, Caltech101, Caltech256, Food101:** class names are directly obtained from the torchvision library.
- **DomainNet-Real:** we normalize directory-based labels and map any non-alphanumeric identifiers via a lookup file.
- **Custom datasets:** if a mapping file (.txt) or explicit class list is provided, the loader automatically uses it; otherwise, class names are inferred from folder names or dataset metadata.

Feature storage. For each dataset, the resulting text features are saved under the same feature cache as image embeddings with shape (C, d), where C is the number of classes and d is the CLIP embedding dimension. Each embedding is ℓ_2-normalized and averaged across all templates, providing stable, semantically grounded class prototypes.

C. Image feature extraction

To enable computationally efficient and reproducible active learning, CCMA operates entirely in a frozen feature space. For all experiments, image embeddings are precomputed using pretrained vision backbones and stored in a shared cache. These embeddings are then used to train a simple linear classifier (student head) without updating the backbone weights.

Backbones. We employ pretrained models from standard vision foundations: DINOv2 [30] for the student and CLIP ViT-L/14 [33] for the teacher. The student model serves as the vision-only encoder producing embeddings $z(x)$, while the teacher's CLIP image encoder produces multimodal-aligned embeddings $\phi(x)$. Both backbones remain frozen throughout all active learning iterations.

Feature extraction and caching. Each dataset split (train/test) is passed through the frozen backbone once to extract image features. Given an image x_i, the feature vector $f_i = \phi(x_i)$ is computed after the global pooling layer and ℓ_2-normalized. All extracted features and their class labels are stored in an HDF5 cache. This allows CCMA to reuse embeddings across rounds, significantly reducing compute time.

Linear classifier training. During active learning, only a linear classification head $W \in \mathbb{R}^{C\times D}$ is trained on top of the frozen student features. This design isolates the acquisition strategy from representation learning, ensuring that improvements stem purely from better sample selection rather than feature fine-tuning.

D. Active learning training on the student model

Architecture. The student model is a simple linear classifier $h_c(x) = W\,\psi_\theta\big(z(x)\big) + b$ trained on top of the frozen DINOv2 [30] features. The classifier includes an optional dropout layer for Monte Carlo sampling and uncertainty estimation. No gradients are propagated to the backbone, making each training round efficient and stable.

Training loop. During each active learning iteration:

1. The student model is trained on the currently labeled pool $\mathcal{L}$ using cross-entropy loss.
2. Dropout-based sampling and conformal calibration yield per-sample predictive sets $\Gamma_S(x)$ and $\Gamma_T(x)$.
3. The disagreement between teacher and student posteriors guides acquisition scoring.
4. Selected samples are labeled and added to $\mathcal{L}$ for the next round.

Prediction modes. The model supports multiple inference modes:

- `"probs"` – outputs softmax probabilities for uncertainty-based scoring.
- `"embed"` – returns intermediate embeddings for feature-space clustering.
- `"grad"` – computes gradient embeddings (e.g., for BADGE-style acquisition).

This flexibility allows fair comparisons with diverse active learning strategies while maintaining a consistent linear classifier backbone.

E. Implementation details of split-conformal calibration

In the main paper (Section 3.3), we introduced two calibrated set predictors (one for the teacher ($m = T$) and one for the student ($m = S$)) based on split-conformal calibration. Here, we provide additional details on the implementation, parameter settings, and rationale for the chosen target set sizes.

Nonconformity scores and calibration. For both predictors, the nonconformity score is defined as

$$a_m(x, c) = -\log p_m(c \mid x),$$

where $p_m(c \mid x)$ denotes the softmax probability assigned to class c by model $m \in \{T, S\}$. A calibration split $\mathcal{C}_{\text{cal}}$ (disjoint from the active pool) is used to compute empirical thresholds q_m such that either the expected set size or the coverage constraint is satisfied.

Size-targeted calibration. We target a desired mean prediction set size s_m by solving for the smallest threshold q_m that satisfies

$$\frac{1}{|\mathcal{C}_{\text{cal}}|} \sum_{(x,y)\in\mathcal{C}_{\text{cal}}} |\{c : a_m(x, c) \le q_m\}| \approx s_m.$$

This threshold q_m is obtained via bisection search on the sorted nonconformity scores, ensuring the average conformal set size matches the target s_m.

Practical thresholds. In all experiments, we fix the target set sizes as:

$$s_S = 5, \qquad s_T = 3.$$

That is, the student predictor $\Gamma_S(x)$ includes the top 5 most probable classes on average, while the teacher predictor $\Gamma_T(x)$ includes the top 3. These choices were determined empirically to balance informativeness and calibration stability: - smaller s_T yields sharper and more confident teacher sets, focusing guidance on the most relevant semantic hypotheses; - slightly larger s_S encourages exploration and helps the student capture broader class uncertainty during early rounds.

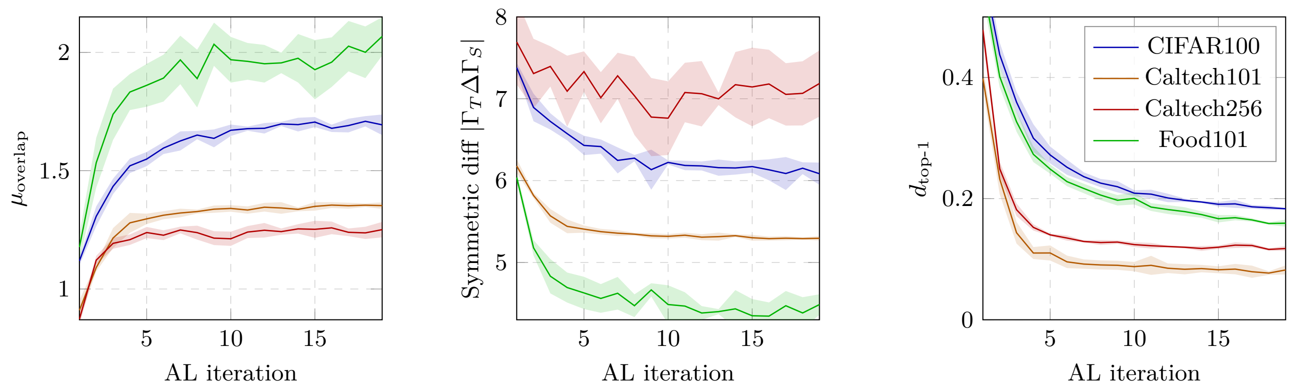
The calibrated thresholds q_m are recomputed at each active learning iteration using the current model's predictions on the calibration split, ensuring adaptivity as the student improves.

Coverage and reliability. Although our setup targets fixed expected set sizes, we verify that the resulting coverage (the fraction of samples whose true label lies in $\Gamma_m(x)$) remains high across rounds: typically 0.90–0.95 for the teacher and 0.85–0.90 for the student. This behavior is consistent with the conformal guarantee that for monotone a_m, quantile-based thresholds yield calibrated coverage independent of data distribution.

In summary, our conformal calibration procedure introduces no additional learnable parameters and incurs negligible computational overhead. The fixed target set sizes s_T=3 and s_S=5 were consistently used across all datasets (CIFAR-100, Food-101, DomainNet-Real, and Caltech variants) and yielded stable, reproducible conformal thresholds for both teacher and student predictors.

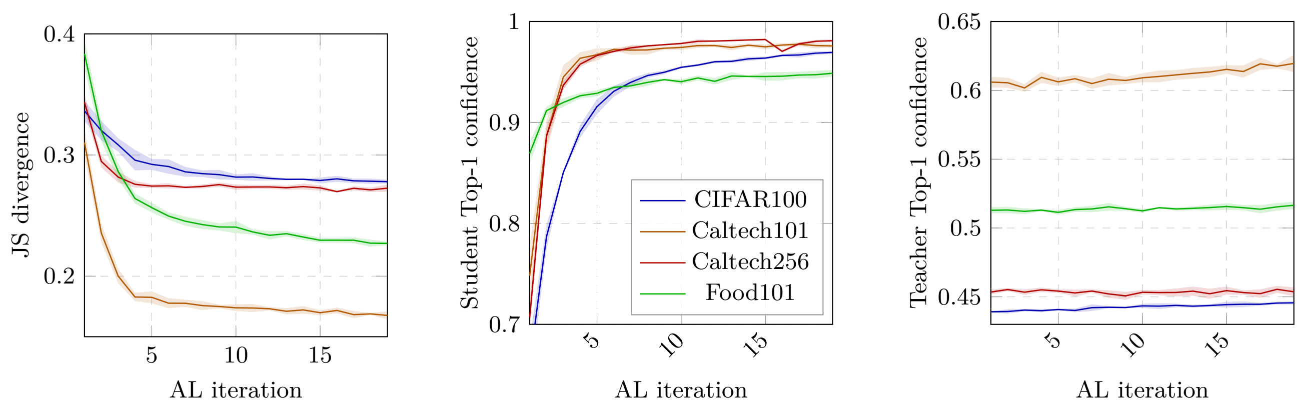
F. Identifying regimes where teacher signals no longer help

A core challenge in multimodal active learning is determining when a VLM teacher provides meaningful



(a) Mean CCMA overlap (b) Mean CCMA symmetric difference (c) Mean fraction Top-1 Disagreement

Figure 8. CCMA diagnostic curves across AL iterations. (a) Overlap between teacher and student conformal sets, measuring the agreement in predicted label supports. Higher overlap indicates stronger cross-modal consistency. (b) Symmetric difference between the two sets, capturing the size of their disagreement region. Larger values reflect a persistent teacher–student mismatch. (c) Top-1 disagreement rate, the probability that the teacher and student assign different most-likely classes. These metrics reveal how the informativeness of teacher–student disagreement evolves: datasets like CIFAR100 and Food101 exhibit sustained mismatch (useful for CCMA), whereas Caltech101/256 align quickly, reducing the value of cross-modal uncertainty and shifting the problem toward coverage-based selection.



(a) JS Divergence (b) Student Top-1 Confidence (c) Teacher Top-1 Confidence

Figure 9. Teacher–student confidence dynamics across AL iterations. (a) JS divergence between teacher and student posteriors restricted to their shared support, quantifying cross-modal disagreement. (b) Student top-1 confidence, showing how quickly the student becomes certain as labeled data accumulates. (c) The teacher's top-1 confidence, which remains relatively stable across iterations since the teacher is frozen. These curves highlight the regimes where CCMA is most effective: datasets like CIFAR100 and Food101 maintain moderate teacher–student disagreement (useful for cross-modal query scoring), whereas Caltech101/256 exhibit rapid confidence saturation and early agreement, reducing the benefit of disagreement-based acquisition.

guidance. To answer this, we track four complementary diagnostics across AL iterations: **(i)** CCMA set overlap $\mu_{\text{overlap}} = |\Gamma_T \cap \Gamma_S|$ (Fig. 8a), **(ii)** symmetric difference $|\Gamma_T \triangle \Gamma_S|$ (Fig. 8b), **(iii)** Top-1 disagreement $d_{\text{top-1}}$ (Fig. 8c), and **(iv)** JS divergence and predictive confidence trends (Fig. 9).

CIFAR100 and Food101. Across early and mid AL rounds, both disagreement and JS divergence decay gradually (Figs. 8c and 9a), indicating that the teacher continues to provide complementary information as the student improves. Overlap steadily increases while symmetric difference decreases (Figs. 8a and 8b), showing convergence in a *content-aware* manner: the student learns semantics that the teacher initially knows better. This persistent teacher–student mismatch keeps multimodal uncertainty informative, enabling CCMA to maintain strong gains throughout the budget.

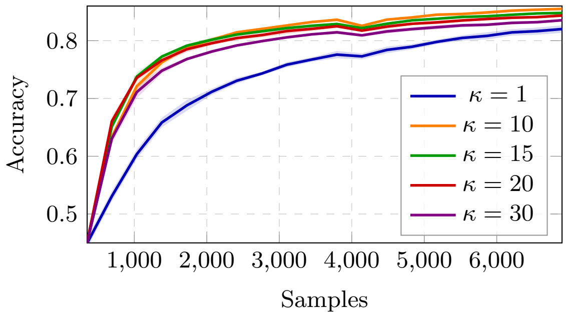


Figure 10. Effect of oversampling factor κ on DomainNetReal

Caltech101 and Caltech256. In contrast, these datasets display a *fast alignment regime*. The student rapidly becomes highly confident (> 0.95 by $\sim$1k labels; Fig. 9b), and the teacher's confidence remains nearly flat (Fig. 9c), suggesting that both models find Caltech classes relatively easy. JS divergence and top-1 disagreement collapse early (Figs. 9a and 8c), meaning the teacher adds almost no new information. Here, the symmetric difference stabilizes at a low level, and the overlap saturates quickly (Figs. 8b and 8a), revealing that selection becomes *coverage-dominated* rather than uncertainty-driven. Accordingly, clustering-based methods outperform CCMA in later rounds.

These diagnostics expose a previously unaddressed insight: *VLM guidance is most valuable in regimes with persistent semantic mismatch.* Once the teacher is effectively "oracle-like," relying on teacher–student disagreement harms selection efficiency. This transforms CCMA into a **decision rule**: use cross-modal conformal uncertainty when divergence remains high, else transition toward pure coverage as mismatch vanishes.

G. Impact of hyperparameter choice

Figure 10 illustrates the sensitivity of CCMA to the oversampling factor κ on the DomainNet-Real dataset. Recall that κ controls how many top-ranked candidates (by disagreement score) are retained before the final diversity-based clustering. Increasing κ enlarges this candidate set, potentially improving selection diversity at the cost of additional computation. We observe that performance improves significantly when moving from $\kappa = 1$ (no oversampling) to $\kappa = 10$, confirming that moderate oversampling effectively stabilizes selection and enhances early learning. However, beyond $\kappa = 10$, accuracy gains saturate and slightly decline, indicating diminishing returns and possible noise amplification in large candidate pools. This trend contrasts with CIFAR100 and Food101, where higher κ values (15–20) yielded more consistent improvements. Overall, $\kappa = 10$ provides the best trade-off between accuracy and efficiency for DomainNet-Real, balancing exploration and stability during sample acquisition.

Variant	AULC	Final@R	Time/rd (s)	AULC/s
V1	0.859	0.915	3.015	0.285
V2	0.828	0.899	0.780	1.061
V3	0.835	0.906	0.512	1.629
V4	0.809	0.896	2.929	0.276
V5	0.793	0.885	0.583	1.360

Table 2. Mean over 5 seeds. Final accuracy at the last round (Final@R), AULC computed over rounds, and AULC per second on the CIFAR100 dataset.

H. Ablation details

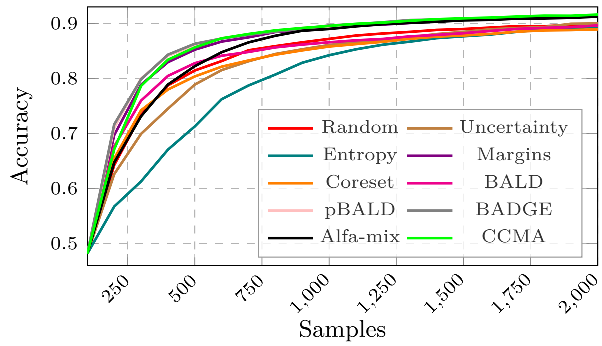
Table 2 reports the average performance across five seeds on CIFAR100, comparing several CCMA variants (V1–V5) that differ in their subpooling or diversity-selection configurations. We evaluate each variant in terms of the *Area Under the Learning Curve* (AULC), final accuracy at the last round (Final@R), mean computation time per active learning round, and overall efficiency measured as AULC per second. As shown, V1 achieves the highest overall accuracy and AULC, indicating the strongest acquisition quality, but requires the longest runtime due to full conformal and clustering operations. Conversely, lightweight variants (e.g., V3 and V5) attain slightly lower accuracy but significantly higher time-normalized efficiency (AULC/s), demonstrating that simplified selection can substantially accelerate active learning without drastic performance loss. These results emphasize that CCMA offers a controllable balance between computational cost and selection effectiveness, enabling adaptation to different practical budgets and latency constraints.

I. All the experiments

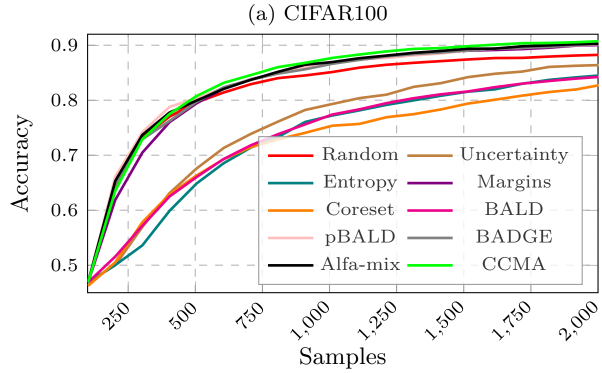
For completeness, we attempted to include all representative active learning baselines in our experiments. However, two recent clustering-based approaches, **TypiClust** [18] and **ProbCover** [48], could not be consistently evaluated across all datasets due to inherent constraints in their selection logic. In addition, we conducted supplementary experiments by varying the teacher–student backbone configurations, specifically using **DINOv2 ViT-L/14** and **DINOv2 ViT-g/14** as students, and **CLIP ViT-L/14** and **CLIP ViT-g/14** as teachers, to assess the robustness of CCMA across architectures. The corresponding results are presented in Fig. 11, 12, 13, and 14.

TypiClust. This method clusters the unlabeled pool in feature space and prioritizes *typical* samples from dense regions for annotation. While effective at early stages with small budgets, TypiClust progressively depletes its set of available unlabeled "typical" samples, as the clustering density sharply decreases with each iteration. Consequently, after several rounds (especially on larger datasets such as CIFAR100, Food101, and DomainNet-Real), the algorithm plateaus early and fails to generate new acquisition batches once all cluster centroids have been annotated. In such cases, TypiClust effectively terminates before reaching the full query budget, preventing a fair comparison in later active learning rounds.

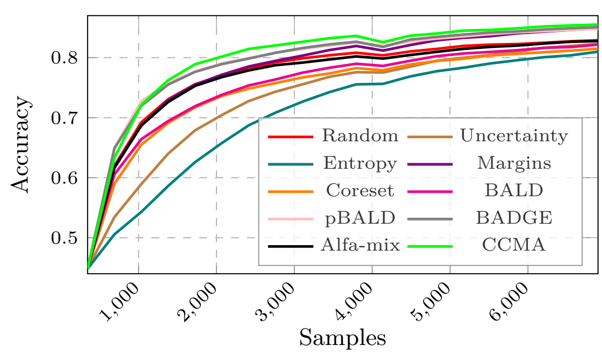
ProbCover. ProbCover selects samples that maximize coverage of high-uncertainty regions under a probabilistic diversity criterion. Its selection mechanism relies on computing representative subsets from the entire unlabeled pool, which becomes unstable or infeasible when the pool is too small or lacks sufficient diversity. This limitation is particularly pronounced in smaller datasets such as Caltech101 and Caltech256, where, after several acquisition rounds, the remaining unlabeled samples no longer satisfy the probabilistic coverage constraints. As a result, the method halts before completing all iterations, leading to missing curves in our plots for these datasets.



(a) CIFAR100

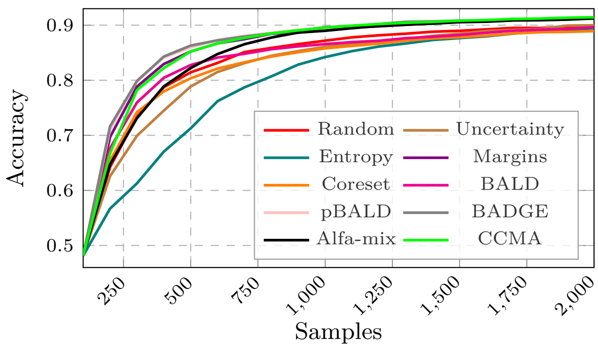


(b) Food101

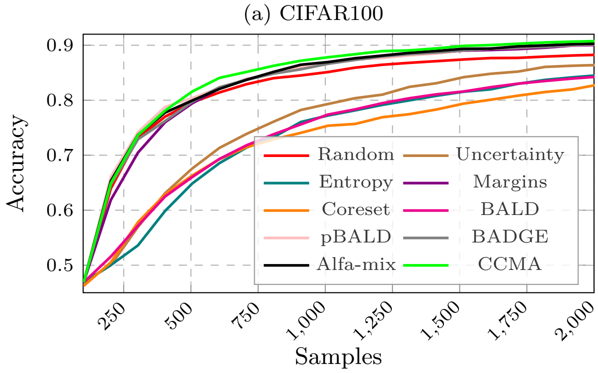


(c) DomainNet-Real

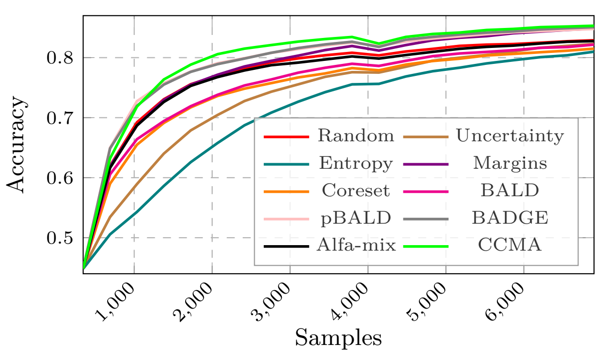
Figure 11. Test mean accuracy over 5 seeds for CCMA with DINOv2 ViT-g14 as a student model and CLIP ViT-L14 as a teacher model with other AL methods on CIFAR100, Food101, and DomainNet-Real datasets.



(a) CIFAR100

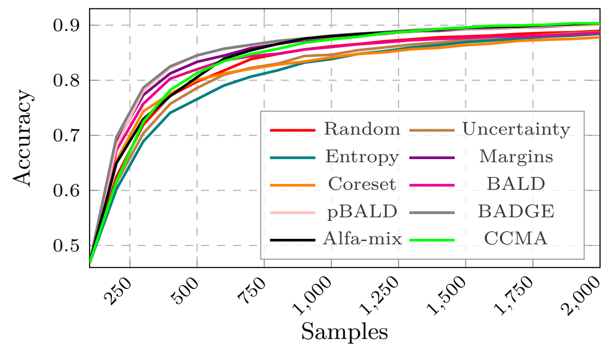


(b) Food101

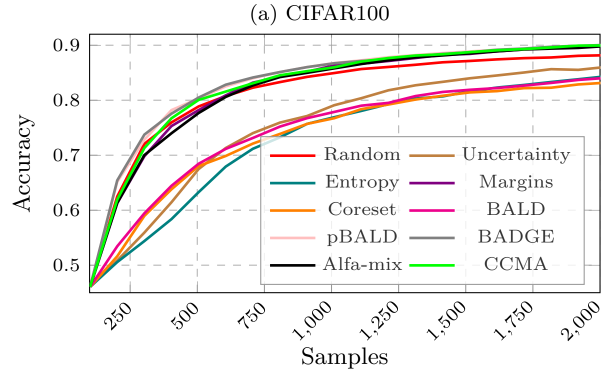


(c) DomainNet-Real

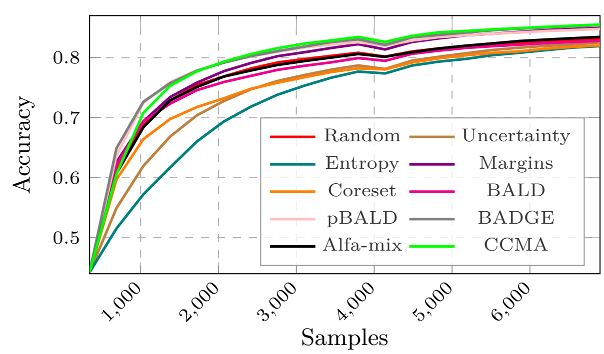
Figure 12. Test mean accuracy over 5 seeds for CCMA with DINOv2 ViT-g14 as a student model and CLIP ViT-g14 as a teacher model with other AL methods on CIFAR100, Food101, and DomainNet-Real datasets.



(a) CIFAR100

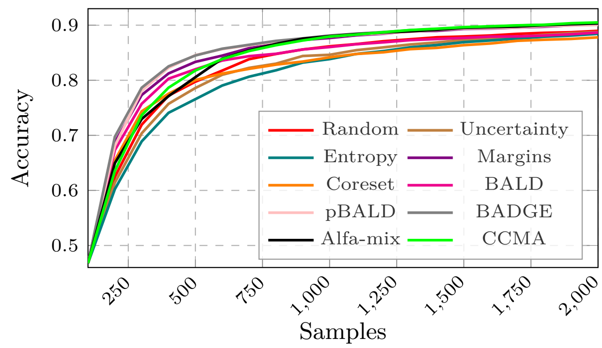


(b) Food101

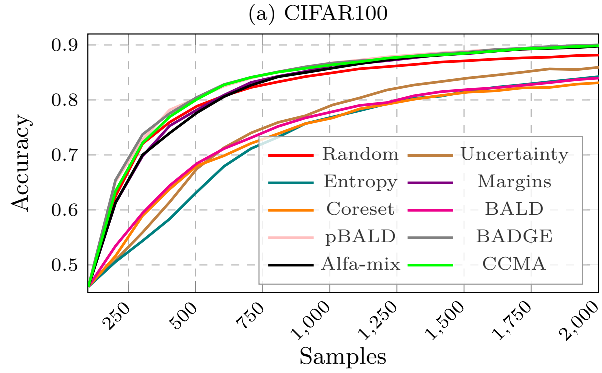


(c) DomainNet-Real

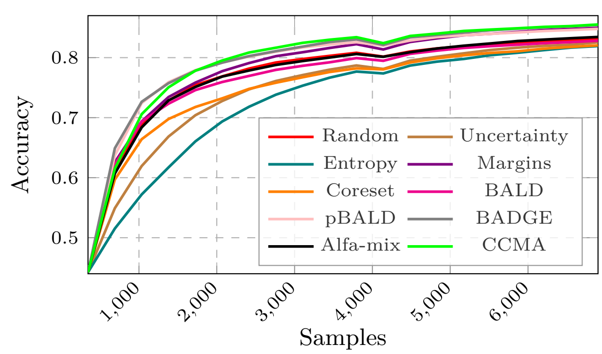
Figure 13. Test mean accuracy over 5 seeds for CCMA with DINOv2 ViT-L14 as a student model and CLIP ViT-L14 as a teacher model with other AL methods on CIFAR100, Food101, and DomainNet-Real datasets.



(a) CIFAR100



(b) Food101



(c) DomainNet-Real

Figure 14. Test mean accuracy over 5 seeds for CCMA with DINOv2 ViT-L14 as a student model and CLIP ViT-g14 as a teacher model with other AL methods on CIFAR100, Food101, and DomainNet-Real datasets.

Table 3. Mean accuracy averaged over 5 runs along with the standard deviation at AL iterations t for datasets CIFAR100 [24], Food101 [7], and DomainNet-Real [32] when utilizing the random initialization with DINOv2 ViT-g14 as the feature extractor. **Bold** values represent the **first-place** mean accuracy at iteration t with the second-place value underlined.

t	Random	Uncertainty	Entropy	Margins	BALD	pBALD	Coreset	BADGE	Alfa-mix	ProbCover	*CCMA (ours)*
						CIFAR100					
1	48.0 ± 2.2	48.0 ± 2.2	48.0 ± 2.2	48.0 ± 2.2	48.0 ± 2.2	48.0 ± 2.2	48.0 ± 2.2	48.0 ± 2.2	48.0 ± 2.2	48.0 ± 2.2	48.0 ± 2.2
2	64.2 ± 2.4	62.5 ± 2.4	56.6 ± 3.0	69.6 ± 0.5	67.3 ± 1.0	71.0 ± 1.3	64.5 ± 2.1	**71.6** ± 1.1	64.7 ± 2.8	70.4 ± 6.6	67.1 ± 1.5
3	73.3 ± 2.1	69.8 ± 1.6	61.2 ± 3.4	78.7 ± 1.0	75.9 ± 0.9	**79.9** ± 0.7	74.2 ± 1.6	79.7 ± 1.0	73.1 ± 2.3	74.4 ± 6.6	78.6 ± 1.3
4	78.7 ± 1.6	74.4 ± 1.8	67.0 ± 1.2	82.6 ± 1.1	80.4 ± 0.5	83.9 ± 0.8	77.9 ± 0.9	**84.1** ± 0.6	78.8 ± 1.4	77.0 ± 5.8	83.4 ± 1.2
5	81.3 ± 0.5	78.7 ± 1.1	71.2 ± 1.1	85.2 ± 1.1	82.7 ± 0.7	85.9 ± 0.5	80.3 ± 0.9	**86.3** ± 0.6	82.2 ± 0.5	78.7 ± 5.8	85.6 ± 0.9
6	83.1 ± 0.4	81.5 ± 0.9	76.2 ± 0.2	86.7 ± 0.9	84.1 ± 0.6	**87.3** ± 0.3	82.1 ± 0.9	87.2 ± 0.3	84.6 ± 1.0	79.5 ± 5.4	**87.3** ± 0.7
7	85.1 ± 0.6	83.1 ± 0.6	78.6 ± 2.3	87.4 ± 0.7	84.8 ± 0.7	88.0 ± 0.6	83.2 ± 1.2	87.8 ± 0.3	86.5 ± 0.8	80.3 ± 5.2	**88.1** ± 0.6
8	85.8 ± 0.7	84.2 ± 0.9	80.7 ± 2.2	87.9 ± 0.7	85.2 ± 0.4	88.4 ± 0.3	84.3 ± 1.0	88.5 ± 0.4	87.7 ± 0.8	81.7 ± 4.6	**88.7** ± 0.6
9	86.4 ± 0.5	85.2 ± 0.6	82.7 ± 1.2	88.6 ± 0.7	86.1 ± 0.3	89.1 ± 0.2	85.0 ± 0.8	89.0 ± 0.2	88.6 ± 0.4	82.0 ± 4.4	**89.2** ± 0.5
10	87.0 ± 0.4	86.0 ± 0.5	84.2 ± 1.1	89.1 ± 0.6	86.5 ± 0.3	89.4 ± 0.3	85.7 ± 0.6	89.5 ± 0.3	88.8 ± 0.2	83.4 ± 3.5	**89.6** ± 0.5
11	87.6 ± 0.7	86.5 ± 0.8	85.2 ± 0.7	89.5 ± 0.4	86.9 ± 0.4	89.7 ± 0.1	86.2 ± 0.6	89.8 ± 0.3	89.4 ± 0.2	84.0 ± 3.6	**89.9** ± 0.5
12	88.0 ± 0.4	87.0 ± 0.9	86.1 ± 0.6	89.7 ± 0.6	87.1 ± 0.6	90.0 ± 0.2	86.6 ± 0.8	90.1 ± 0.2	89.8 ± 0.2	84.7 ± 3.2	**90.2** ± 0.3
13	88.4 ± 0.4	87.4 ± 0.9	86.6 ± 0.6	90.0 ± 0.3	87.6 ± 0.5	90.1 ± 0.3	87.1 ± 1.0	**90.5** ± 0.2	90.0 ± 0.3	85.2 ± 2.9	**90.5** ± 0.4
14	88.7 ± 0.2	88.0 ± 0.6	87.3 ± 0.4	90.3 ± 0.2	87.9 ± 0.5	90.3 ± 0.1	87.5 ± 0.9	90.6 ± 0.0	90.2 ± 0.2	85.5 ± 2.8	**90.8** ± 0.4
15	88.9 ± 0.3	88.4 ± 0.6	87.6 ± 0.3	90.7 ± 0.1	88.2 ± 0.2	90.5 ± 0.3	87.8 ± 0.8	90.7 ± 0.2	90.5 ± 0.2	85.9 ± 2.7	**90.9** ± 0.4
16	89.2 ± 0.2	88.9 ± 0.8	87.8 ± 0.5	90.7 ± 0.2	88.4 ± 0.3	90.6 ± 0.2	88.3 ± 0.8	90.8 ± 0.1	90.6 ± 0.2	86.3 ± 2.6	**91.1** ± 0.4
17	89.5 ± 0.1	89.1 ± 0.8	88.4 ± 0.4	91.1 ± 0.2	89.0 ± 0.2	90.9 ± 0.2	88.5 ± 0.7	91.1 ± 0.1	90.8 ± 0.2	86.7 ± 2.3	**91.3** ± 0.3
18	89.4 ± 0.1	89.3 ± 0.6	88.9 ± 0.5	91.0 ± 0.2	89.1 ± 0.3	91.0 ± 0.1	88.7 ± 0.7	91.1 ± 0.1	90.9 ± 0.1	86.9 ± 2.1	**91.4** ± 0.2
19	89.7 ± 0.2	89.9 ± 0.6	89.1 ± 0.5	91.1 ± 0.2	89.3 ± 0.3	91.2 ± 0.1	88.8 ± 0.8	91.2 ± 0.1	91.0 ± 0.1	87.5 ± 1.9	**91.4** ± 0.3
20	89.8 ± 0.2	89.9 ± 0.5	89.4 ± 0.1	91.2 ± 0.0	89.6 ± 0.3	91.2 ± 0.2	88.9 ± 0.6	91.3 ± 0.1	91.2 ± 0.2	88.0 ± 1.8	**91.6** ± 0.2
						Food101					
1	46.8 ± 1.9	46.8 ± 1.9	46.8 ± 1.9	46.8 ± 1.9	46.8 ± 1.9	46.8 ± 1.9	46.8 ± 1.9	46.8 ± 1.9	46.8 ± 1.9	46.8 ± 1.9	46.8 ± 1.9
2	63.9 ± 2.1	49.9 ± 3.2	50.0 ± 1.8	61.8 ± 2.7	50.8 ± 2.3	66.0 ± 2.3	50.5 ± 3.0	64.6 ± 1.7	65.3 ± 2.1	**76.9** ± 0.7	63.5 ± 1.0
3	73.1 ± 0.3	56.8 ± 3.4	53.6 ± 2.3	70.4 ± 2.1	57.1 ± 1.8	74.1 ± 3.0	57.8 ± 2.8	72.8 ± 1.4	73.6 ± 2.1	**80.2** ± 1.1	72.9 ± 1.1
4	77.0 ± 1.1	63.0 ± 3.2	58.6 ± 1.5	75.4 ± 2.1	62.2 ± 2.4	78.1 ± 2.1	62.8 ± 3.7	76.2 ± 1.8	77.7 ± 1.6	**82.0** ± 0.9	78.1 ± 0.9
5	79.6 ± 0.9	67.5 ± 2.1	64.7 ± 2.6	79.4 ± 1.8	66.0 ± 2.1	80.2 ± 1.9	66.3 ± 3.1	79.7 ± 1.2	79.8 ± 1.4	**83.7** ± 0.9	80.8 ± 1.0
6	81.3 ± 1.0	71.3 ± 2.1	68.5 ± 3.0	81.1 ± 1.6	69.3 ± 1.7	82.2 ± 2.0	69.2 ± 3.3	82.3 ± 1.4	82.0 ± 1.4	**84.7** ± 0.8	83.2 ± 0.6
7	82.8 ± 1.0	73.7 ± 2.0	71.3 ± 2.8	83.7 ± 1.4	71.8 ± 1.8	83.6 ± 0.9	71.4 ± 3.4	83.7 ± 1.6	83.6 ± 1.1	**85.7** ± 0.6	84.6 ± 0.0
8	83.8 ± 0.3	76.1 ± 1.8	73.0 ± 2.4	84.7 ± 0.7	73.8 ± 2.3	85.0 ± 0.8	72.8 ± 2.9	84.8 ± 0.9	85.1 ± 1.1	86.0 ± 0.5	**86.1** ± 0.5
9	84.5 ± 0.4	78.2 ± 1.9	76.0 ± 2.0	85.7 ± 0.6	75.6 ± 2.1	86.1 ± 0.8	74.0 ± 2.9	85.6 ± 0.9	86.4 ± 0.4	86.3 ± 0.6	**86.8** ± 0.6
10	85.1 ± 0.1	79.3 ± 1.8	77.2 ± 2.1	86.6 ± 0.9	77.3 ± 2.6	86.6 ± 0.9	75.3 ± 2.3	86.5 ± 0.4	86.8 ± 0.4	86.7 ± 0.4	**87.7** ± 0.7
11	85.8 ± 0.1	80.3 ± 1.9	78.1 ± 1.4	87.4 ± 0.9	78.3 ± 2.0	87.5 ± 0.5	75.6 ± 2.2	87.3 ± 0.7	87.6 ± 0.6	87.3 ± 0.3	**88.3** ± 0.7
12	86.4 ± 0.3	81.0 ± 2.1	79.1 ± 2.1	87.7 ± 0.8	79.4 ± 1.6	87.6 ± 0.4	76.8 ± 2.1	87.8 ± 0.7	88.0 ± 0.5	87.4 ± 0.6	**88.9** ± 0.3
13	86.8 ± 0.2	82.4 ± 1.7	79.8 ± 2.2	88.3 ± 1.0	80.3 ± 1.6	88.1 ± 0.5	77.4 ± 1.7	88.3 ± 0.6	88.6 ± 0.1	87.8 ± 0.6	**89.3** ± 0.0
14	87.1 ± 0.3	83.0 ± 1.3	80.7 ± 1.4	88.5 ± 0.9	81.0 ± 1.9	88.4 ± 0.3	78.3 ± 1.8	88.6 ± 0.4	88.9 ± 0.3	88.3 ± 0.4	**89.5** ± 0.3
15	87.3 ± 0.1	84.1 ± 1.4	81.5 ± 1.3	89.0 ± 0.8	81.5 ± 2.1	89.0 ± 0.2	79.3 ± 1.8	89.0 ± 0.4	89.2 ± 0.1	88.4 ± 0.5	**89.8** ± 0.2
16	87.4 ± 0.4	84.8 ± 1.5	81.9 ± 1.5	89.2 ± 0.8	82.3 ± 1.7	89.3 ± 0.3	80.0 ± 1.4	89.3 ± 0.3	89.1 ± 0.2	88.4 ± 0.3	**90.1** ± 0.2
17	87.7 ± 0.2	85.2 ± 1.0	82.9 ± 1.1	89.2 ± 0.8	83.0 ± 1.2	89.5 ± 0.3	80.8 ± 1.4	89.5 ± 0.5	89.8 ± 0.1	88.7 ± 0.5	**90.3** ± 0.2
18	87.9 ± 0.3	86.0 ± 0.8	83.7 ± 0.8	89.5 ± 0.7	83.4 ± 1.1	89.7 ± 0.3	81.4 ± 1.7	89.7 ± 0.3	89.9 ± 0.0	89.0 ± 0.7	**90.4** ± 0.1
19	88.1 ± 0.3	86.2 ± 0.8	84.1 ± 0.8	89.9 ± 0.2	83.9 ± 1.4	89.9 ± 0.2	81.9 ± 1.3	90.0 ± 0.2	90.1 ± 0.2	89.2 ± 0.4	**90.5** ± 0.1
20	88.2 ± 0.3	86.4 ± 1.0	84.5 ± 0.9	90.0 ± 0.1	84.3 ± 1.1	90.1 ± 0.3	82.8 ± 1.1	90.2 ± 0.3	90.2 ± 0.1	88.7 ± 0.5	**90.8** ± 0.2
						DomainNet-Real					
1	44.7 ± 0.8	44.7 ± 0.8	44.7 ± 0.8	44.7 ± 0.8	44.7 ± 0.8	44.7 ± 0.8	44.7 ± 0.8	44.7 ± 0.8	44.7 ± 0.8	44.7 ± 0.8	44.7 ± 0.8
2	61.8 ± 0.9	53.4 ± 1.3	50.4 ± 1.6	61.9 ± 0.8	60.2 ± 0.7	64.9 ± 1.2	58.9 ± 1.2	64.8 ± 1.0	61.6 ± 0.5	**72.6** ± 0.8	65.1 ± 0.7
3	69.2 ± 0.5	58.9 ± 1.8	54.2 ± 2.5	68.6 ± 0.5	66.4 ± 0.6	72.6 ± 0.8	65.4 ± 0.9	71.9 ± 0.6	68.6 ± 0.6	**75.1** ± 0.3	72.1 ± 0.8
4	73.0 ± 0.7	64.0 ± 1.1	58.6 ± 2.7	72.9 ± 0.5	69.4 ± 0.5	75.7 ± 0.6	69.1 ± 1.1	75.5 ± 0.6	72.6 ± 0.3	76.3 ± 0.6	**77.7** ± 0.5
5	75.5 ± 0.6	67.8 ± 1.0	62.5 ± 2.4	75.4 ± 0.6	71.8 ± 0.3	77.6 ± 0.5	71.7 ± 0.5	77.6 ± 0.4	75.2 ± 0.2	77.7 ± 0.6	**78.9** ± 0.3
6	76.9 ± 0.2	70.4 ± 0.7	65.7 ± 1.7	77.1 ± 0.7	73.8 ± 0.2	79.0 ± 0.3	73.6 ± 0.2	78.8 ± 0.6	76.7 ± 0.5	78.2 ± 0.6	**80.2** ± 0.4
7	78.1 ± 0.3	74.3 ± 0.6	68.7 ± 1.3	78.4 ± 0.5	75.3 ± 0.5	79.9 ± 0.4	74.8 ± 0.3	79.8 ± 0.2	77.8 ± 0.4	78.7 ± 0.8	**81.5** ± 0.2
8	79.2 ± 0.2	74.3 ± 0.6	70.8 ± 0.8	79.4 ± 0.5	76.3 ± 0.6	80.8 ± 0.2	75.7 ± 0.4	80.8 ± 0.1	78.7 ± 0.4	78.9 ± 0.8	**82.0** ± 0.2
9	79.8 ± 0.2	75.5 ± 0.3	72.6 ± 0.7	80.3 ± 0.5	77.4 ± 0.5	81.4 ± 0.2	76.7 ± 0.4	81.6 ± 0.1	79.1 ± 0.3	79.4 ± 0.7	**82.6** ± 0.2
10	80.2 ± 0.0	76.8 ± 0.3	74.2 ± 0.4	81.2 ± 0.4	78.3 ± 0.7	82.0 ± 0.1	77.3 ± 0.4	82.2 ± 0.1	79.6 ± 0.2	79.3 ± 0.6	**83.2** ± 0.2
11	80.8 ± 0.1	77.5 ± 0.2	75.5 ± 0.4	81.8 ± 0.2	78.8 ± 0.6	82.5 ± 0.3	78.2 ± 0.4	82.6 ± 0.2	80.2 ± 0.2	79.4 ± 0.7	**83.6** ± 0.3
12	80.3 ± 0.0	77.5 ± 0.3	75.6 ± 0.3	81.1 ± 0.2	78.6 ± 0.5	81.9 ± 0.3	77.8 ± 0.5	81.7 ± 0.2	79.8 ± 0.3	79.3 ± 0.5	**82.6** ± 0.2
13	81.0 ± 0.2	78.4 ± 0.4	76.8 ± 0.5	82.0 ± 0.3	79.4 ± 0.4	82.6 ± 0.3	78.8 ± 0.4	83.0 ± 0.1	80.4 ± 0.1	79.6 ± 0.5	**83.6** ± 0.1
14	81.4 ± 0.2	79.4 ± 0.3	77.7 ± 0.5	82.8 ± 0.2	80.2 ± 0.5	83.1 ± 0.3	79.4 ± 0.5	83.4 ± 0.1	80.8 ± 0.2	79.4 ± 0.7	**84.0** ± 0.2
15	81.9 ± 0.2	79.9 ± 0.3	78.3 ± 0.4	83.3 ± 0.1	80.6 ± 0.6	83.5 ± 0.3	79.8 ± 0.6	83.8 ± 0.2	81.4 ± 0.4	79.4 ± 0.5	**84.5** ± 0.2
16	82.1 ± 0.3	80.4 ± 0.4	79.0 ± 0.4	83.5 ± 0.2	80.9 ± 0.4	83.9 ± 0.2	80.3 ± 0.5	84.2 ± 0.2	81.7 ± 0.5	79.5 ± 0.5	**84.6** ± 0.2
17	82.2 ± 0.1	81.0 ± 0.3	79.5 ± 0.4	84.0 ± 0.0	81.2 ± 0.5	84.2 ± 0.3	80.6 ± 0.5	84.3 ± 0.1	82.0 ± 0.5	79.4 ± 0.4	**84.9** ± 0.2
18	82.4 ± 0.2	81.6 ± 0.3	80.0 ± 0.4	84.3 ± 0.1	81.6 ± 0.3	84.4 ± 0.2	80.9 ± 0.5	84.6 ± 0.2	82.3 ± 0.6	79.6 ± 0.6	**85.2** ± 0.2
19	82.7 ± 0.1	81.9 ± 0.2	80.3 ± 0.2	84.6 ± 0.1	81.7 ± 0.3	84.6 ± 0.2	81.1 ± 0.7	84.9 ± 0.1	82.6 ± 0.6	79.7 ± 0.4	**85.4** ± 0.1
20	82.8 ± 0.1	82.2 ± 0.2	80.9 ± 0.3	84.8 ± 0.0	82.1 ± 0.3	84.7 ± 0.2	81.4 ± 0.6	85.0 ± 0.1	82.7 ± 0.7	79.7 ± 0.5	**85.5** ± 0.1

Table 4. Test accuracy for our method (*CCMA*) with other AL strategies when utilizing the random initialization with DINOv2 ViT-g14 as the feature extractor. We show AL iterations t for datasets Caltech101 and Caltech256 when utilizing the random initialization. **Bold** values represent the **first-place** mean accuracy at iteration t with the $\underline{\text{second-place}}$ value $\underline{\text{underlined}}$.

t	Random	Uncertainty	Entropy	Margins	BALD	pBALD	Coreset	BADGE	Alfa-mix	Typiclust	*CCMA (ours)*
						Caltech-256					
1	57.4 ± 1.5	57.4 ± 1.5	57.4 ± 1.5	57.4 ± 1.5	57.4 ± 1.5	57.4 ± 1.5	57.4 ± 1.5	57.4 ± 1.5	57.4 ± 1.5	57.4 ± 1.5	57.4 ± 1.5
2	76.4 ± 1.4	70.0 ± 1.6	66.3 ± 1.6	78.9 ± 1.0	78.1 ± 0.6	$\underline{82.9}$ ± 1.1	81.5 ± 1.1	**83.9** ± 0.8	75.8 ± 1.0	75.3 ± 1.8	82.2 ± 1.0
3	83.8 ± 1.4	77.7 ± 2.8	72.0 ± 1.5	86.0 ± 0.9	85.1 ± 0.5	$\underline{90.0}$ ± 0.3	87.0 ± 0.5	89.7 ± 0.6	83.7 ± 0.7	84.4 ± 1.0	**90.8** ± 0.8
4	87.9 ± 1.0	83.8 ± 1.3	76.5 ± 1.9	89.6 ± 0.5	87.8 ± 0.7	$\underline{92.3}$ ± 0.5	89.3 ± 0.7	92.0 ± 0.3	88.5 ± 0.9	88.4 ± 1.3	**93.1** ± 0.4
5	90.4 ± 0.7	87.2 ± 1.3	80.8 ± 1.9	92.1 ± 0.5	90.1 ± 0.7	93.5 ± 0.2	90.7 ± 0.7	$\underline{93.8}$ ± 0.4	90.6 ± 0.7	91.3 ± 1.5	**94.1** ± 0.2
6	92.0 ± 0.3	89.9 ± 1.4	84.4 ± 1.4	93.8 ± 0.3	91.8 ± 0.8	94.5 ± 0.2	92.0 ± 0.6	$\underline{94.7}$ ± 0.5	92.0 ± 0.4	92.3 ± 0.5	**94.8** ± 0.1
7	92.7 ± 0.4	92.6 ± 1.0	87.5 ± 1.2	94.8 ± 0.4	93.1 ± 0.7	$\underline{95.2}$ ± 0.3	92.7 ± 0.7	**95.6** ± 0.4	92.9 ± 0.4	93.1 ± 0.2	$\underline{95.2}$ ± 0.1
8	93.4 ± 0.3	93.9 ± 0.8	89.9 ± 1.3	95.5 ± 0.1	94.1 ± 0.4	$\underline{95.8}$ ± 0.2	93.2 ± 0.8	**96.2** ± 0.3	93.8 ± 0.5	93.5 ± 0.3	95.4 ± 0.1
9	94.0 ± 0.3	95.1 ± 0.5	91.7 ± 1.0	$\underline{96.3}$ ± 0.1	94.8 ± 0.5	96.2 ± 0.2	93.9 ± 0.6	**97.0** ± 0.3	94.6 ± 0.7	94.1 ± 0.3	95.8 ± 0.0
10	94.4 ± 0.2	96.1 ± 0.3	93.4 ± 0.7	$\underline{96.9}$ ± 0.2	95.4 ± 0.2	96.6 ± 0.2	94.4 ± 0.5	**97.3** ± 0.3	95.5 ± 0.9	94.3 ± 0.3	95.9 ± 0.0
11	94.6 ± 0.1	96.7 ± 0.2	94.3 ± 0.8	$\underline{97.1}$ ± 0.1	95.9 ± 0.3	97.0 ± 0.3	94.9 ± 0.3	**97.7** ± 0.3	95.7 ± 0.9	94.5 ± 0.3	96.1 ± 0.0
12	94.8 ± 0.1	97.2 ± 0.3	95.5 ± 0.7	$\underline{97.5}$ ± 0.2	96.5 ± 0.2	97.3 ± 0.1	95.1 ± 0.3	**98.0** ± 0.2	96.0 ± 0.9	94.7 ± 0.2	96.3 ± 0.1
13	95.0 ± 0.2	97.6 ± 0.3	96.5 ± 0.6	$\underline{97.9}$ ± 0.2	97.0 ± 0.2	97.6 ± 0.1	95.6 ± 0.2	**98.2** ± 0.1	96.5 ± 1.1	94.9 ± 0.2	96.4 ± 0.0
14	95.1 ± 0.2	$\underline{98.0}$ ± 0.1	97.0 ± 0.6	$\underline{98.0}$ ± 0.2	97.3 ± 0.2	$\underline{98.0}$ ± 0.1	95.7 ± 0.3	**98.4** ± 0.1	96.7 ± 1.0	95.0 ± 0.1	96.6 ± 0.0
15	95.3 ± 0.1	$\underline{98.0}$ ± 0.2	97.1 ± 0.4	**98.1** ± 0.1	97.5 ± 0.3	97.9 ± 0.1	96.0 ± 0.2	97.8 ± 0.2	97.0 ± 0.8	95.1 ± 0.0	96.7 ± 0.0
16	94.8 ± 0.1	97.4 ± 0.2	97.2 ± 0.3	$\underline{97.7}$ ± 0.1	97.4 ± 0.3	97.6 ± 0.1	95.5 ± 0.2	**97.9** ± 0.2	96.6 ± 0.5	94.8 ± 0.9	96.3 ± 0.1
17	95.3 ± 0.1	$\underline{98.5}$ ± 0.2	98.0 ± 0.3	**98.8** ± 0.1	98.0 ± 0.1	98.3 ± 0.0	96.2 ± 0.1	**98.8** ± 0.1	97.4 ± 0.5	95.4 ± 0.1	96.7 ± 0.0
18	95.4 ± 0.2	98.7 ± 0.2	98.4 ± 0.3	$\underline{99.0}$ ± 0.1	98.4 ± 0.2	98.6 ± 0.0	96.3 ± 0.2	**99.1** ± 0.2	97.8 ± 0.5	95.5 ± 0.1	96.9 ± 0.1
19	95.6 ± 0.1	$\underline{98.8}$ ± 0.2	98.7 ± 0.3	**99.2** ± 0.1	98.7 ± 0.1	98.7 ± 0.0	96.5 ± 0.3	**99.2** ± 0.1	98.0 ± 0.4	95.6 ± 0.0	97.1 ± 0.0
20	95.8 ± 0.1	$\underline{99.1}$ ± 0.1	98.9 ± 0.3	**99.3** ± 0.0	98.9 ± 0.0	98.8 ± 0.1	96.8 ± 0.1	**99.3** ± 0.0	98.3 ± 0.4	95.7 ± 0.1	97.2 ± 0.1
						Caltech-101					
1	61.5 ± 2.0	61.5 ± 2.0	61.5 ± 2.0	61.5 ± 2.0	61.5 ± 2.0	61.5 ± 2.0	61.5 ± 2.0	61.5 ± 2.0	61.5 ± 2.0	61.5 ± 2.0	61.5 ± 2.0
2	77.0 ± 1.4	75.1 ± 3.1	71.0 ± 2.2	82.3 ± 2.5	82.2 ± 2.0	85.2 ± 1.0	**88.6** ± 1.4	$\underline{86.7}$ ± 1.1	76.9 ± 2.7	76.4 ± 2.0	75.7 ± 1.7
3	84.9 ± 2.2	83.5 ± 1.2	79.7 ± 1.5	88.5 ± 0.8	88.5 ± 1.3	**90.6** ± 0.5	$\underline{90.1}$ ± 0.5	**90.6** ± 0.8	84.1 ± 1.8	83.7 ± 1.2	82.2 ± 0.8
4	89.4 ± 0.9	87.8 ± 0.6	85.8 ± 1.5	90.4 ± 0.8	90.3 ± 0.6	$\underline{92.1}$ ± 0.7	90.8 ± 0.7	**92.2** ± 0.7	88.0 ± 1.5	88.3 ± 0.4	86.1 ± 0.6
5	90.9 ± 0.3	89.6 ± 0.5	88.3 ± 1.4	92.5 ± 0.8	91.0 ± 0.5	$\underline{93.6}$ ± 0.7	90.9 ± 0.7	**94.1** ± 0.8	89.8 ± 1.5	90.4 ± 1.1	90.0 ± 0.7
6	92.5 ± 0.5	91.5 ± 0.6	89.8 ± 0.8	94.1 ± 0.7	91.9 ± 0.7	$\underline{94.7}$ ± 0.7	91.4 ± 0.6	**95.2** ± 0.5	91.1 ± 0.7	91.7 ± 0.6	91.8 ± 0.9
7	93.6 ± 0.7	93.2 ± 0.4	91.2 ± 0.6	94.9 ± 0.6	93.6 ± 0.6	$\underline{95.5}$ ± 0.7	91.5 ± 0.6	**95.8** ± 0.7	92.5 ± 0.5	92.6 ± 0.9	92.8 ± 0.9
8	94.6 ± 0.7	94.5 ± 0.5	92.6 ± 1.1	95.5 ± 0.9	94.7 ± 0.4	$\underline{96.0}$ ± 0.7	91.6 ± 0.8	**96.3** ± 0.8	92.7 ± 0.6	93.4 ± 0.4	94.8 ± 0.1
9	95.0 ± 0.3	95.5 ± 0.5	94.2 ± 1.1	95.9 ± 1.1	95.9 ± 0.5	$\underline{96.5}$ ± 0.7	92.2 ± 0.8	**96.6** ± 0.6	93.3 ± 0.8	93.9 ± 0.3	93.8 ± 0.5
10	95.3 ± 0.3	96.1 ± 0.5	95.1 ± 0.7	96.3 ± 1.0	96.7 ± 0.7	$\underline{96.8}$ ± 0.8	92.7 ± 0.8	**96.9** ± 0.7	94.0 ± 1.2	94.4 ± 0.6	94.5 ± 0.5
11	95.5 ± 0.3	96.3 ± 0.7	95.8 ± 0.8	96.6 ± 1.0	$\underline{97.0}$ ± 0.6	$\underline{97.0}$ ± 0.5	92.7 ± 0.9	**97.2** ± 0.7	95.0 ± 1.7	94.8 ± 0.5	95.1 ± 0.4
12	95.6 ± 0.4	96.7 ± 0.7	96.4 ± 0.8	$\underline{97.0}$ ± 0.9	**97.4** ± 0.6	**97.4** ± 0.6	93.1 ± 1.0	**97.4** ± 0.6	95.9 ± 1.5	94.9 ± 0.5	95.2 ± 0.6
13	95.8 ± 0.4	96.9 ± 0.6	96.8 ± 0.6	97.2 ± 0.9	**97.5** ± 0.6	**97.5** ± 0.6	94.0 ± 0.9	$\underline{97.4}$ ± 0.8	96.6 ± 1.2	95.3 ± 0.4	95.6 ± 0.3
14	96.0 ± 0.3	97.0 ± 0.2	97.1 ± 0.4	97.4 ± 0.8	**97.9** ± 0.4	$\underline{97.6}$ ± 0.4	94.1 ± 0.6	97.5 ± 0.7	97.4 ± 0.4	95.7 ± 0.2	95.7 ± 0.3
15	96.1 ± 0.4	97.3 ± 0.5	97.1 ± 0.3	97.5 ± 0.7	**98.0** ± 0.4	**98.0** ± 0.6	94.3 ± 0.6	$\underline{97.8}$ ± 0.8	**98.0** ± 0.3	95.7 ± 0.1	95.9 ± 0.2
16	96.2 ± 0.2	97.2 ± 0.5	97.2 ± 0.3	97.7 ± 0.6	**98.1** ± 0.3	**98.1** ± 0.6	94.4 ± 0.5	$\underline{97.8}$ ± 0.8	**98.1** ± 0.3	95.8 ± 0.1	96.6 ± 0.2
17	96.3 ± 0.4	97.6 ± 0.5	97.3 ± 0.3	97.9 ± 0.5	**98.4** ± 0.3	$\underline{98.3}$ ± 0.6	94.7 ± 0.7	98.0 ± 0.8	**98.4** ± 0.1	96.1 ± 0.1	96.4 ± 0.2
18	96.6 ± 0.2	98.0 ± 0.8	97.5 ± 0.2	98.2 ± 0.5	$\underline{98.4}$ ± 0.5	$\underline{98.4}$ ± 0.7	94.9 ± 0.6	98.2 ± 0.7	**98.5** ± 0.0	96.3 ± 0.2	96.5 ± 0.2
19	96.7 ± 0.1	98.2 ± 0.6	97.6 ± 0.2	98.4 ± 0.3	$\underline{98.5}$ ± 0.5	$\underline{98.5}$ ± 0.8	95.4 ± 0.7	$\underline{98.5}$ ± 0.3	**98.6** ± 0.1	96.5 ± 0.1	96.5 ± 0.1
20	96.8 ± 0.3	98.3 ± 0.6	98.0 ± 0.4	98.4 ± 0.5	$\underline{98.5}$ ± 0.5	$\underline{98.5}$ ± 0.7	95.6 ± 0.5	$\underline{98.5}$ ± 0.4	**98.6** ± 0.0	96.6 ± 0.1	96.7 ± 0.2