Conformal Cross-Modal Active Learning

Huy Hoang Nguyen[1], Cédric Jung[1,2], Shirin Salehi[3], Tobias Glück[1], Anke Schmeink[3], Andreas Kugi[1,2]

[1] AIT Austrian Institute of Technology [2] Automation and Control Institute, Technical University of Vienna

[3] Chair of Information Theory and Data Analytics (INDA), RWTH Aachen University

{huy-hoang.nguyen,cedric.jung,tobias.glueck,andreas.kugi}@ait.ac.at,

{shirin.salehi,anke.schmeink}@inda.rwth-aachen.de

Abstract

Foundation models for vision have transformed visual recognition with powerful pretrained representations and strong zero-shot capabilities, yet their potential for data-efficient learning remains largely untapped. Active Learning (AL) aims to minimize annotation costs by strategically selecting the most informative samples for labeling, but existing methods largely overlook the rich multimodal knowledge embedded in modern vision–language models (VLMs). We introduce ***Conformal Cross-Modal Acquisition (CCMA)****, a novel AL framework that bridges vision and language modalities through a teacher–student architecture. CCMA employs a pretrained VLM as a teacher to provide semantically grounded uncertainty estimates, conformally calibrated to guide sample selection for a vision-only student model. By integrating multimodal conformal scoring with diversity-aware selection strategies, CCMA achieves superior data efficiency across multiple benchmarks. Our approach consistently outperforms state-of-the-art AL baselines, demonstrating clear advantages over methods relying solely on uncertainty or diversity metrics.*

1. Introduction

In recent years, artificial intelligence has undergone a paradigm shift with the rise of foundation models, such as DALL-E [34], GPT-3 [8], Dinov2 [30], trained on broad data at scale. While these models provide powerful, transferable visual representations, their development requires massive amounts of curated data and computation resources. This challenge is especially pronounced in classification tasks, where large annotated datasets remain essential for high accuracy [11]. To alleviate annotation and training costs, Active Learning (AL) has emerged as a compelling framework that aims to reduce annotation requirements by selecting the most informative samples for labeling [21, 26, 38, 39].

Pretrained visual features from foundation models have recently improved AL pipelines [9, 16, 44], but most existing approaches remain *vision-only*. Vision Language Models (VLMs) [6], such as CLIP [33], offer an untapped opportunity: their text–image alignment captures high-level class semantics, suggesting that they could provide more informative signals for sample selection than standard visual features alone. Preliminary attempts to use VLMs for AL focus mostly on prompt tuning [4, 5], leaving unexplored how to extract or quantify uncertainty from multimodal representations.

A key challenge arises: VLM outputs are often miscalibrated, domain-dependent, and not directly comparable to task-specific classifier probabilities, limiting their direct use as uncertainty oracles. Conformal Prediction (CP) [1, 29, 42, 43, 45] offers an appealing solution by providing distribution-free, per-sample uncertainty sets that remain valid regardless of model architecture or miscalibration. However, existing conformal AL methods [22, 28] operate strictly within a single modality and do not leverage cross-modal semantic structure. Likewise, existing VLM-based AL does not use CP to fuse information from different models.

This gap motivates our research question: *Can we incorporate the semantic structure of VLMs into an active learning acquisition function using distribution-free conformal calibration?* We answer affirmatively by developing **Conformal Cross-Modal Acquisition (CCMA)**, a novel AL framework that bridges vision and language modalities through calibrated uncertainty estimation. CCMA employs a pretrained VLM as a teacher to generate semantically grounded prediction sets, which are conformally calibrated to provide

distribution-free uncertainty estimates for guiding a vision-only student model. By integrating cross-modal conformal scoring with diversity-aware selection strategies, CCMA achieves superior data efficiency across multiple benchmarks.

Our main contributions are as follows:

1. We propose a **teacher–student conformal scoring mechanism** that aligns vision-only predictions from a student model with text–image guidance from a pretrained VLM teacher. By constructing conformal prediction sets calibrated on held-out data, CCMA provides *distribution-free, per-sample uncertainty* that is robust across datasets and architectures.
2. We introduce a **selective subpooling strategy** based on clustering in CLIP [33] feature space, which preserves geometric diversity while substantially reducing the number of candidates to be scored. Combined with an **uncertainty-weighted coverage** objective, CCMA achieves an effective trade-off between scalability and informativeness, enabling efficient active selection without accuracy degradation.
3. We conduct extensive experiments on multiple image classification benchmarks, showing that CCMA consistently outperforms state-of-the-art active learning baselines across diverse domains and modalities.
4. We provide a detailed analysis of the role of VLMs in active learning, revealing that CCMA excels when meaningful teacher–student discrepancies exist, while performance saturates once the teacher approaches oracle accuracy—transitioning the challenge from uncertainty estimation to coverage optimization.

2. Related Works

2.1. Active Learning

In pool-based AL [14, 19, 41] and classification tasks, the AL problem can be defined as follows: The whole dataset at first presents a small labeled dataset part named $\mathcal{L} = \{(x_j, y_j)\}_{j=1}^{M}$ and a larger unlabeled part named $\mathcal{U} = \{x_i\}_{i=1}^{N}$, where $M \ll N$, $y_i \in \{0, 1\}$ is the class label of x_i for binary classification, or $y_i \in \{1, \ldots, C\}$ for multi-class classification with C classes. The process involves selecting instances from the unlabeled dataset $\mathcal{U}$ in a greedy manner, guided by a set of informativeness metrics called *acquisition functions*. In each iteration t, a batch $\mathcal{D}_t^*$ of size B from $\mathcal{U}$ is selected based on the learned model $\mathcal{M}$ and an acquisition function $\mathcal{A}(x, \mathcal{M})$, and queries their labels from the oracle. Data samples can be selected according to their acquisition score by $\mathcal{D}_t^* = \operatorname{argmax}^B_{x \in \mathcal{U}} \mathcal{A}(x, \mathcal{M})$, where the superscript B indicates selection of the top B points. In general, to reduce computational costs, a subpool $\mathcal{D}_{\text{pool}}$ is drawn from the unlabeled dataset $\mathcal{U}$, on which the acquisition function will be computed: $\mathcal{D}_t^* = \operatorname{argmax}^B_{x \in \mathcal{D}_{\text{pool}}} \mathcal{A}(x, \mathcal{M})$.

Acquisition functions. Uncertainty-based methods query unlabeled instances where the model is most uncertain. *Entropy* [46] measures total predictive uncertainty by selecting samples with the highest entropy, while *Margins* [36] examines the gap between the top two class probabilities, selecting those with the smallest margin. Another *Uncertainty* baseline [25] selects samples with the lowest maximum predicted probability. Beyond these point-estimate approaches, Bayesian active learning by disagreement (*BALD*) [14] targets examples that maximize mutual information between predictions and the parameter posterior. *PowerBALD* [23] extends this by accounting for correlations among queried samples, thereby reducing the redundancy inherent in top-B selection strategies. Most recently, *UHerding* [3] maximizes an uncertainty-weighted coverage objective by using calibrated uncertainties and a shrinking kernel to balance diversity at low budgets with pure uncertainty at high budgets.

The choice between uncertainty-based and representativity/diversity-based acquisition functions reflects an exploration–exploitation trade-off, motivating hybrid approaches that combine or alternate between both [20, 47]. Mixup-based methods, such as *Alfa-mix* [31], interpolate unlabeled samples to create synthetic queries, whereas our approach operates directly in feature space and uses probabilistic set constructions to guide informative selection. Representative-based methods aim to cover the data distribution. Classical approaches such as the k-center greedy method [12] and *Coreset* [40] minimize distances to cluster centers. Similarly, *Typiclust* [18] queries typical points within clusters, while *ProbCover* [48] avoids outliers to improve representativeness. Leveraging clustering within a semantically meaningful feature space that is obtained via self-supervised learning further promotes diverse sampling.

2.2. Conformal prediction

Conformal prediction (CP) [45] provides distribution-free, finite-sample uncertainty quantification by producing *prediction sets* that contain the true label with user-specified coverage under the exchangeability assumption [42]. It has been explored in AL to quantify model uncertainty for regression [28] and classi-

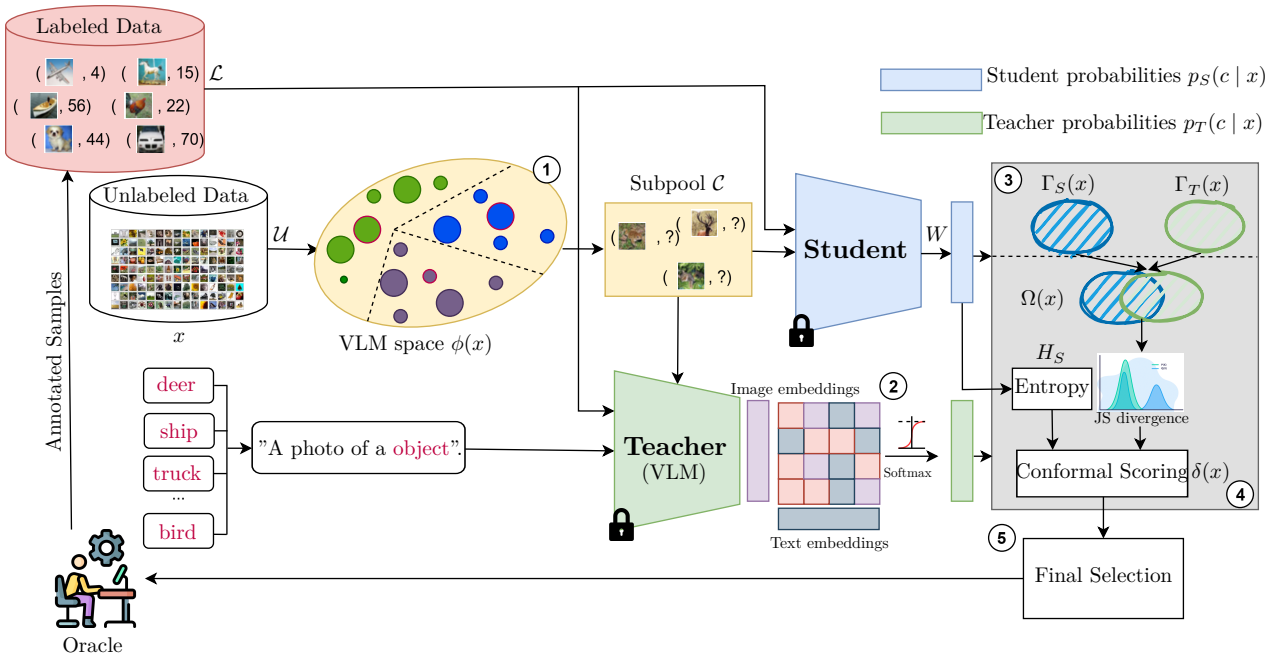


Figure 1. **Overview of our proposed AL framework CCMA for image classification by exploring conformal prediction with multimodal uncertainty and diversity for efficient data acquisition**. Given labeled data $\mathcal{L}$ and an unlabeled pool $\mathcal{U}$, a frozen vision encoder (student) and a frozen VLM (teacher) serve as feature extractors, while a simple linear classifier is trained on the student features. **(1)** A selective subpool is formed via CLIP-space clustering. **(2–3)** Student and teacher posteriors p_S, p_T are calibrated into conformal sets Γ_S, Γ_T. **(4)** Multimodal disagreement $\Omega(x)$ combines entropy and Jensen-Shannon (JS) divergence for uncertainty scoring $\delta(x)$. **(5)** Top-ranked samples are oversampled and clustered with uncertainty-weighted coverage for the final diverse selection.

fication [27] tasks. One approach ranks samples by conformal uncertainty, selecting those with the smallest p-values or the largest nonconformity (CoPAL [22] and related CPAL variants [28]). Another approach aims to reduce annotation costs by querying candidate class sets that are guaranteed to include the true label, prioritizing examples with small yet reliable sets [17]. However, these methods are typically *unimodal*, assigning each example a scalar based on a single model's conformity or set size, while addressing diversity only heuristically.

2.3. Active Learning in VLMs

Foundation models [6], including vision–language models (VLMs), are trained on large unlabeled or noisy data, learning representations that enable strong zero- and few-shot performance. These properties make them natural candidates for active learning (AL), which seeks to maximize labeling efficiency. Although the two paradigms can complement each other [5], their integration remains mainly underexplored [16]. Yet, applying conventional AL frameworks to pretrained VLMs can sometimes degrade performance [4, 49], motivating the development of AL strategies tailored to VLMs. For instance, [37] combines calibrated entropy with self- and neighbor-aware uncertainty to produce more reliable selection scores, narrowing the zero-shot–supervised gap of VLMs. Unlike their work, which applies AL for prompt tuning, we leverage cross-modal knowledge from VLM teachers to guide sample selection for vision-only students.

Concluding, while AL seeks to minimize annotation cost by selectively querying the most informative samples, its effectiveness is often hindered by unreliable uncertainty estimates from purely vision-based models. Recent VLMs provide semantically rich, cross-modal representations that can guide AL toward more meaningful and transferable sample selection. In this paper, we address this limitation by introducing a conformal prediction framework that bridges the uncertainty gap between visual and textual modalities, thereby enhancing both the efficiency and robustness of AL.

3. Our method: Conformal Cross-Model Acquisition (CCMA)

We propose a multi-modal conformal acquisition function that integrates diversity sampling with uncertainty

estimation, using a student-teacher disagreement score. The approach follows a five-stage process, as depicted in Fig. 1.

3.1. Diverse subpool selection

We adopt a *compute-aware* candidate selection strategy by compressing the unlabeled pool $\mathcal{U}$ into a smaller candidate set $\mathcal{C}$ of size $|\mathcal{C}| \ll |\mathcal{U}|$. To construct $\mathcal{C}$, we cluster the VLM image embeddings $\phi(x) : x \in \mathcal{U}$ in ϕ-space (introduced in Sec. 3.2) and choose one or a few representative points from each cluster using k-Means.

We then apply the subsequent steps of CCMA exclusively to $\mathcal{C}$. This approach reduces the per-round scoring cost from $O(|\mathcal{U}|B)$ to $O(|\mathcal{C}|B)$ with $|\mathcal{C}| \ll |\mathcal{U}|$, while preserving coverage of the pool. Empirically, this diversity-selected subpool consistently outperforms an equally sized random subpool, yielding higher accuracy at low budgets and maintaining robustness without incurring the high computation of full-pool selection.

3.2. Two predictors: student and teacher

For a given sample image x, we use image embeddings from pretrained VLM $\phi(x) \in \mathbb{R}^d$ and text prototypes $\{t_c\}_{c=1}^C \subset \mathbb{R}^d$, all ℓ_2-normalized. The VLM teacher produces logits $\ell_T(c \mid x) = \phi(x)^\top t_c/\tau$, where τ is the temperature parameter. The posterior is then obtained via the softmax function:

$$p_T(c \mid x) = \frac{\exp(\ell_T(c \mid x))}{\sum_{c'=1}^C \exp(\ell_T(c' \mid x))}. \tag{1}$$

The student classifier f_S operates on a separate backbone feature $z(x) \in \mathbb{R}^D$ (e.g., extracted from DINOv2) and outputs class probabilities $p_S(c \mid x)$. The student consists of a feature adapter $\psi_\theta(\cdot)$ and a linear classification head with weight matrix $W \in \mathbb{R}^{C\times D}$ with bias $b \in \mathbb{R}^C$, $h_c(x) = W\,\psi_\theta\big(z(x)\big) + b$, yielding:

$$p_S(c \mid x) = \frac{\exp\big(h_c(x)\big)}{\sum_{c'=1}^C \exp\big(h_{c'}(x)\big)}. \tag{2}$$

We train the parameters θ, W, and b using cross-entropy loss on the labeled set and evaluate the best checkpoint to obtain $\{p_S(c \mid x)\}_{c=1}^C$ for all x in the candidate subpool.

3.3. Two calibrated set predictors (split conformal)

For conformal prediction calibration, we define the nonconformity score as $a_m(x, c) = -\log p_m(c \mid x)$, where $p_m(c \mid x)$ denotes the predicted probability of class c given input x, and $m \in \{T, S\}$ refers to the teacher T and student S. In a calibration split $\mathcal{C}_{\text{cal}} \subseteq \mathcal{D}$, we determine thresholds q_m for either the target expected set size s_m or the marginal coverage $1 - \alpha_m$.

For size-targeted calibration, we find q_m by bisection such that

$$\frac{1}{|\mathcal{C}_{\text{cal}}|} \sum_{(x,y)\in\mathcal{C}_{\text{cal}}} |\{c : a_m(x, c) \le q_m\}| \approx s_m. \tag{3}$$

For coverage-targeted calibration, we set q_m as the empirical $(1 - \alpha_m)$-quantile of nonconformity scores $\{a_m(x, y) : (x, y) \in \mathcal{C}_{\text{cal}}\}$, guaranteeing split-conformal marginal coverage $\ge 1 - \alpha_m$, where $\alpha_m \in [0, 1]$ represents the tolerated error rate. The set-valued predictors are then

$$\Gamma_m(x) = \{c \in [C] : a_m(x, c) \le q_m\}. \tag{4}$$

We never force-add labels to $\Gamma_m(x)$; if sets are too small/large, we resolve q_m to meet (3). Because the calibration is split-conformal and applied independently to both modalities, this procedure is distribution-free and does not assume that the VLM teacher is well-calibrated or domain-aligned. Unlike prior AL methods that rely on raw VLM logits or treat VLMs as oracle predictors [37], our calibration guarantees valid finite-sample coverage for both teacher and student, making the cross-modal guidance robust even under severe teacher miscalibration or distribution shift.

3.4. Cross-modal disagreement scoring

Given the conformal label sets $\Gamma_S(x)$ and $\Gamma_T(x)$, their union support is defined as $\Omega(x) = \Gamma_S(x) \cup \Gamma_T(x) \subseteq \{1, \ldots, C\}$. For any posterior $p_m(\cdot \mid x) \in \Delta^{C-1}$, with $m \in \{T, S\}$, the distribution is renormalized over $\Omega(x)$:

$$p_m^\Omega(c \mid x) = \frac{p_m(c \mid x)\,\mathbb{1}\{c \in \Omega(x)\}}{\sum_{c'\in\Omega(x)} p_m(c' \mid x)}, \tag{5}$$

with indicator function $\mathbb{1}$. Once the renormalized posteriors p_T^Ω and p_S^Ω are obtained, the Jensen–Shannon (JS) divergence between them is computed as:

$$\begin{aligned}\mathrm{JS}(p_T^\Omega \| p_S^\Omega) = {} & \tfrac{1}{2}\,\mathrm{KL}\Big(p_T^\Omega \,\Big\|\, \tfrac{p_T^\Omega + p_S^\Omega}{2}\Big) \\ & + \tfrac{1}{2}\,\mathrm{KL}\Big(p_S^\Omega \,\Big\|\, \tfrac{p_T^\Omega + p_S^\Omega}{2}\Big),\end{aligned} \tag{6}$$

where $\mathrm{KL}(p \,\|\, r)$ is the Kullback–Leibler divergence between distributions p and r. To dynamically balance the contributions of the student and teacher predictions, we introduce a parameter-free confidence gate. First, we define the top-1 confidence scores for the student and teacher models as $\mathrm{conf}_S(x) = \max_c p_S(c \mid x)$, and $\mathrm{conf}_T(x) = \max_c p_T(c \mid x)$, respectively. Using these confidences, the confidence gate weight is computed as

$$w_{\text{js}}(x) = \frac{\mathrm{conf}_T(x)}{\mathrm{conf}_T(x) + \mathrm{conf}_S(x) + \varepsilon} \in [0, 1], \tag{7}$$

where $\varepsilon > 0$ is added for numerical stability. Intuitively, $w_{\text{js}}(x)$ increases when the teacher is more confident than the student, thereby giving the teacher a greater influence on the final prediction, and vice versa. In early AL rounds, the teacher typically exhibits higher confidence and thus contributes more strongly to the score, whereas in later rounds, the student naturally takes over as its predictions sharpen. This adaptivity arises without additional hyperparameters and avoids committing to either a fully teacher-driven or student-driven rule, leading to the final score:

$$\delta(x) = w_{\text{js}}(x)\ \mathrm{JS}\Big(p_S^{\Omega}\ \Big\|\ p_T^{\Omega}\Big) + \big(1 - w_{\text{js}}(x)\big)\ H_S\big(y \mid x\big), \tag{8}$$

where $H_S\big(y \mid x\big)$ is the entropy of the student's predictions:

$$H_S\big(y \mid x\big) \;=\; -\sum_{c=1}^{C} p_S(c \mid x)\ \log p_S(c \mid x). \tag{9}$$

3.5. Uncertainty-weighted coverage selection

To construct a query batch of size B, we first define a candidate set

$$S_{\kappa B} = \{\, x_i \in \mathcal{U} \mid r_i \le \kappa B \,\}, \quad r_i = \mathrm{rank}_{\downarrow}(\delta(x_i))$$

where $\delta(x_i)$ denotes the disagreement-based uncertainty score and $\kappa \ge 1$ the oversampling factor. Hence, $S_{\kappa B}$ contains the κB most uncertain samples, from which the final batch $S \subseteq S_{\kappa B}$, $|S| = B$, is selected by maximizing the *uncertainty-weighted coverage*:

$$F(S) \;=\; \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \delta(u)\ \max_{s \in S} k_\sigma\big(\phi(u), \phi(s)\big), \tag{10}$$

where k_σ is a similarity kernel on the embedding space $\phi(\cdot)$. The factor κ trades off efficiency and accuracy: a larger κ improves diversity and coverage at the cost of longer query time.

4. Experiments

4.1. Experimental Setups

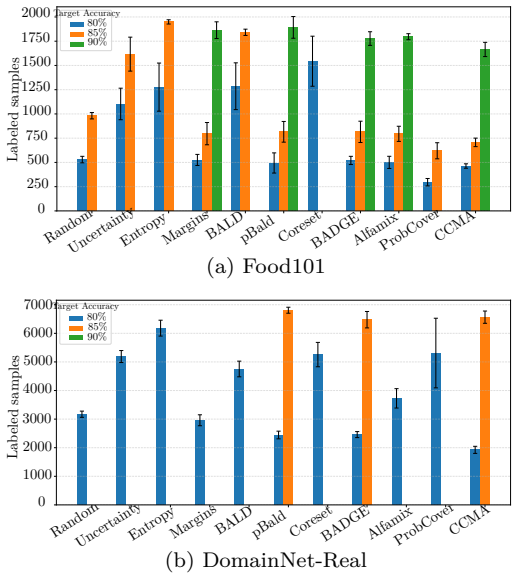
Datasets. We evaluate our CCMA's performance against a suite of state-of-the-art AL methods across several benchmark datasets: CIFAR100 [24], Food101 [7], and DomainNet-Real [32], Caltech101 [13], Caltech256 [15] (see Appendix A).

Implementation details. We employ a frozen CLIP [33] ViT-L/14 [10] model as the **teacher**, where the text encoder provides class-wise prototypes for each downstream category using standard prompts such as "A photo of a [CLS]." (see Appendix B). For each sample, logits are computed via temperature-scaled cosine similarity between image and text embeddings as described in Section 3.2, with a fixed temperature $\tau = 0.01$ (CLIP default) and $\tau = 0.03$ for ablations. The **student** is a frozen vision-only backbone from DINOv2 [30] followed by a linear classification head.

For active learning, we follow the training protocol of [16], running $t = 20$ iterations with five seeds {1, 10, 100, 1000, 10000}, and report the mean accuracy averaged over seeds. Unless otherwise stated, we fix the oversampling factor $\kappa = 20$, target set sizes $s_T = 3$ and $s_S = 5$, and investigate the sensitivity to κ in Section 4.4. Each query round acquires B samples, e.g., $B = 100$ for CIFAR100. Linear heads are trained using AdamW with a learning rate of 10^{-2}, weight decay of 10^{-2}, and dropout rate $\rho = 0.75$.

Baselines. We benchmark results using our method with *eleven* AL baselines: random sampling (Random), uncertainty-based (Entropy, Margins, Uncertainty, BALD, BADGE, PowerBALD denoted by pBALD, Alfa-mix), representation-based (Coreset, Typiclust, ProbCover).

4.2. Experimental Results



(a) Food101

(b) DomainNet-Real

Figure 2. Labels required to reach target accuracies of 80% (blue), 85% (orange), and 90% (green) on Food101 and DomainNet-Real. Lower values indicate higher label efficiency. CCMA consistently reaches each accuracy threshold with fewer labeled samples than uncertainty- and coverage-based baselines, demonstrating improved sample efficiency across both datasets, especially in low-budget regimes.

Result analysis. Across all datasets, CCMA competitively outperforms existing active learning strategies in both early- and late-stage acquisition, demonstrating its effectiveness in leveraging cross-modal uncertainty for sample selection (Table 1). On **CIFAR100**, CCMA achieves the highest final accuracy of 91.6%, surpassing the strongest baseline (BADGE) by +0.3% and maintaining a clear advantage throughout all acquisition rounds. **Food101** and **DomainNet-Real** show similar trends, with CCMA reaching 90.8% and 85.5%, respectively, outperforming all competing uncertainty- and diversity-based methods, including BALD, pBALD, and Coreset. The gains are slightly improved in the early iterations ($t \leq 8$), where most baselines suffer from unstable uncertainty estimates, whereas CCMA benefits from the calibrated teacher–student disagreement, which provides more reliable per-sample confidence. Moreover, the improvements persist even in later rounds, indicating that CCMA does not merely focus on high-entropy regions but maintains semantic coverage through conformal calibration and diversity-aware selection. Notably, our conformal scoring adaptively balances the influence of teacher and student confidence, allowing the model to rely more on the student as its predictions become more reliable in later rounds, while still leveraging teacher guidance in early uncertain stages. This dynamic interplay enables CCMA to achieve a better balance between exploration (via subpool diversity) and exploitation (via multimodal uncertainty), leading to consistent gains across acquisition cycles.

Label efficiency analysis. In Fig. 2a, on **Food101**, CCMA and ProbCover are the only methods able to reach 85% accuracy with fewer than 750 labeled samples. ProbCover performs competitively in the very low-budget regime due to its coverage-driven selection, which encourages early diversity and helps the model learn coarse class boundaries with minimal supervision. However, as the labeling budget increases, its lack of calibrated uncertainty limits further improvement. In contrast, CCMA maintains strong performance across all budget levels, the only method that achieves 90% accuracy with fewer than 1.8K labeled samples, demonstrating its ability to exploit both uncertainty and diversity in a calibrated multimodal manner.

On the more challenging **DomainNet-Real** benchmark, which features significant domain variability and cross-category visual shifts, CCMA stands out as the only method capable of rapidly reaching 80% accuracy with fewer than 2K labeled samples, as shown in Fig. 2b. In contrast, other active learning baselines require substantially more annotations to achieve comparable performance. At higher accuracy thresholds, where many AL methods fail to progress beyond 85%, CCMA, Badge, and PowerBALD remain the only approaches able to reach this level, albeit with larger budgets exceeding 6K samples. These results demonstrate that CCMA provides a favorable balance between early-stage label efficiency and sustained learning capacity, maintaining competitive performance even as the annotation budget increases.

4.3. When is the student better than the teacher in AL?

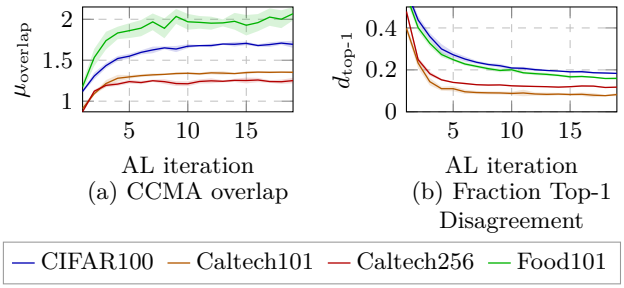


Figure 3. CCMA diagnostics in the overlap and the fraction Top-1 disagreement between teacher and student.

Modern active learning typically assumes a single learner, whereas CCMA introduces a teacher–student interaction in which a pretrained VLM teacher guides a vision-only student through conformal uncertainty calibration. But does this guidance always help, and under which conditions can the student surpass the teacher?

To address this, we analyze benchmarks using two diagnostics: mean CCMA overlap ($\mu_{\text{overlap}} = \frac{1}{|\mathcal{U}|}\sum_{x\in\mathcal{U}} |\Gamma_T(x) \cap \Gamma_S(x)|$) and Top-1 disagreement fraction ($d_{\text{top-1}} = \frac{1}{|\mathcal{U}|}\sum_{x\in\mathcal{U}} \mathbb{1}[\arg\max p_T(x) \neq \arg\max p_S(x)]$) shown in Fig. 3. On CIFAR100 and Food-101, the growing overlap and gradual decay of disagreement indicate that the teacher and student maintain complementary uncertainties across several rounds. CCMA capitalizes on this sustained mismatch, enabling more label-efficient learning and improved accuracy. In contrast, on Caltech101 and Caltech256, both overlap and disagreement flatten early, indicating that the student rapidly aligns with the teacher. Once this stage is reached, the teacher provides only a little new information, and the AL selection process becomes limited. As a result, uncertainty-based methods (e.g., BALD or BADGE) gain advantages in later rounds. This trend is reflected in the accuracy curves (Fig. 4), where CCMA outperforms Random and other baselines, and achieves the highest accuracy on

Table 1. 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.

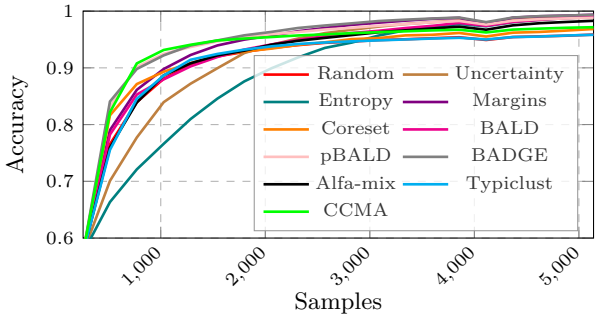
t	Random	Uncertainty	Entropy	Margins	BALD	pBALD	Coreset	BADGE	Alfa-mix	ProbCover	*CCMA (ours)*
						CIFAR100					
1	48.0 ± 2.2	48.0 ± 2.2	48.0 ± 2.2	48.0 ± 2.2	48.0 ± 2.2	48.0 ± 2.2	48.0 ± 2.2	48.0 ± 2.2	48.0 ± 2.2	48.0 ± 2.2	48.0 ± 2.2
4	78.7 ± 1.6	74.4 ± 1.8	67.0 ± 1.2	82.6 ± 1.1	80.4 ± 0.5	$\underline{83.9} \pm 0.8$	77.9 ± 0.9	$\mathbf{84.1} \pm 0.6$	78.8 ± 1.4	77.0 ± 5.8	83.4 ± 1.2
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	$\underline{88.5} \pm 0.4$	87.7 ± 0.8	81.7 ± 4.6	$\mathbf{88.7} \pm 0.6$
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	$\underline{90.8} \pm 0.1$	90.6 ± 0.2	86.3 ± 2.6	$\mathbf{91.1} \pm 0.4$
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	$\underline{91.3} \pm 0.1$	91.2 ± 0.2	88.0 ± 1.8	$\mathbf{91.6} \pm 0.2$
						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
4	77.0 ± 1.1	63.0 ± 3.2	58.6 ± 1.5	75.4 ± 2.1	62.2 ± 2.4	$\underline{78.1} \pm 2.1$	62.8 ± 3.7	76.2 ± 1.8	77.7 ± 1.6	$\mathbf{82.0} \pm 0.9$	$\underline{78.1} \pm 0.9$
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	$\underline{86.0} \pm 0.5$	$\mathbf{86.1} \pm 0.5$
16	87.4 ± 0.4	84.8 ± 1.5	81.9 ± 1.5	89.2 ± 0.8	82.3 ± 1.7	$\underline{89.3} \pm 0.3$	80.0 ± 1.4	$\underline{89.3} \pm 0.3$	89.1 ± 0.2	88.4 ± 0.3	$\mathbf{90.1} \pm 0.2$
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	$\underline{90.2} \pm 0.3$	$\underline{90.2} \pm 0.1$	88.7 ± 0.5	$\mathbf{90.8} \pm 0.2$
						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
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	$\underline{76.3} \pm 0.6$	$\mathbf{77.7} \pm 0.5$
8	79.2 ± 0.2	74.3 ± 0.6	70.8 ± 0.8	79.4 ± 0.5	76.3 ± 0.6	$\underline{80.8} \pm 0.2$	75.7 ± 0.4	$\underline{80.8} \pm 0.1$	78.7 ± 0.4	78.9 ± 0.8	$\mathbf{82.0} \pm 0.2$
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	$\underline{84.2} \pm 0.2$	81.7 ± 0.5	79.5 ± 0.5	$\mathbf{84.6} \pm 0.2$
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	$\underline{85.0} \pm 0.1$	82.7 ± 0.7	79.7 ± 0.5	$\mathbf{85.5} \pm 0.1$

Caltech256 under low-budget settings (< 2K samples) before plateauing as teacher–student disagreement collapses. In contrast, on Caltech101, CCMA remains competitive but only slightly exceeds Coreset and TypiClust, consistent with its coverage-limited nature. Overall, these findings reveal that CCMA excels when meaningful teacher–student discrepancy persists—providing semantically grounded uncertainty signals that guide efficient exploration. Additional diagnostics, including JS divergence and confidence trends, are provided in the supplemental material for completeness (see Appendix F).

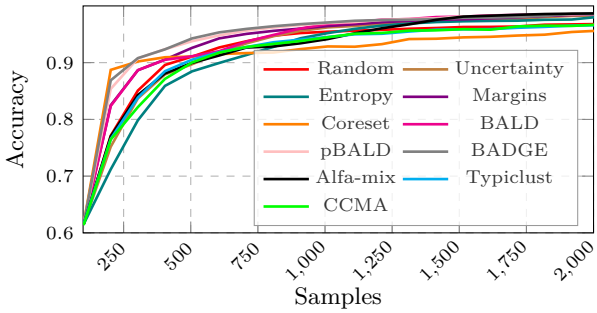
4.4. Impact of hyperparameter choice

The experiments on Food101 and CIFAR100, shown in Fig. 5, reveal that increasing κ from 1 to 20 substantially enhances accuracy, especially when the sample size is small. Beyond $\kappa = 20$, however, the performance gain plateaus, as higher values (e.g., 30) produce nearly identical accuracy curves. This suggests that $\kappa = 20$ effectively captures the benefits of oversampling, with larger values offering negligible additional improvements.

The student's performance is directly tied to the clarity of the teacher's signals. Fig. 6 demonstrates that a lower temperature parameter ($\tau = 0.01$) provides stronger, more informative supervision, enabling the student to achieve higher accuracy with fewer samples. Conversely, a higher τ ($\tau = 0.30$) dilutes the teacher's guidance. However, as labeled size increases, the student trained without teacher guidance eventually reaches the same high-τ teacher's performance. The student ultimately benefits from strong initial guidance, but may outgrow the teacher's utility as the labeled set grows. This motivates CCMA's hybrid rule,

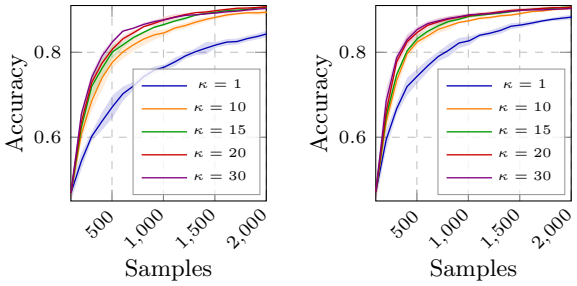


(a) Caltech256



(b) Caltech101

Figure 4. Test mean accuracy over 5 seeds for CCMA with other AL methods on Caltech256 and Caltech101 datasets.



(a) Food101 (b) CIFAR100

Figure 5. Effect of oversampling factor κ.

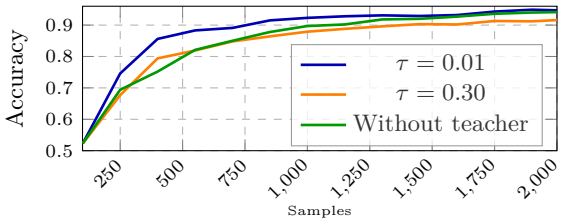


Figure 6. Motivation for investigating the impact of the teacher model in CCMA method. We report the CCMA's accuracy in three different regimes when the teacher is confident ($\tau = 0.01$), weak ($\tau = 0.30$), and disabled on the Food101 benchmark.

which uses teacher–student disagreement early and student entropy once the student becomes more reliable.

4.5. Ablation study

On CIFAR100, Food101, and DomainNet-Real, teacher–student mismatch is initially informative, and CCMA exploits it to achieve the best accuracy and label efficiency. Ablations show both subpool and final diversity matter, while a parameter-free confidence gate preserves performance without dataset-specific tuning. Query-time overhead is minimal, making CCMA a practical, data-aware active learner that knows when to trust the teacher and when to trust itself.

Setup. We evaluate five query variants on CIFAR100 (5 seeds; mean): V1 *(ours)*—selective subpool + final diversity; V2—no subpool + final diversity; V3—random subpool + final diversity; V4—selective subpool, no diversity; V5—no subpool, no diversity. We report the metric Area Under Learning Curve (AULC) [35] over rounds, along with the mean query time/round as shown in Fig. 7.

Across all variants, the component that most directly lifts accuracy is the *final diversity* stage. Removing only diversity (V4) depresses AULC from **0.859** to **0.809**, indicating that disagreement scoring alone is not sufficient to prevent redundancy in selected batches. By contrast, retaining diversity while altering the scope of scoring primarily affects *efficiency*. Scoring the full pool with diversity (V2) is markedly faster than our default (V1), but the curated subpool in V1 yields a clear accuracy margin: **+0.031** AULC, at the cost of longer query time (3.02 s). A large *random* subpool with diversity (V3) provides a particularly attractive trade-off, achieving **0.835** AULC at only **0.51 s**, which is roughly **6**× faster than V1 for a modest AULC loss of **2.4%**. The ablations without diversity (V4, V5) consistently underperform their diversified counterparts (V2, V3), reinforcing that cross-modal disagreement must be paired with within-batch coverage to convert informative uncertainty into label efficiency. Overall, the evidence supports the view that *diversity provides the accuracy gains* while *curated subpooling controls scale*: our full method (V1) achieves the highest accuracy, while V3 provides the best option under strict time budgets.

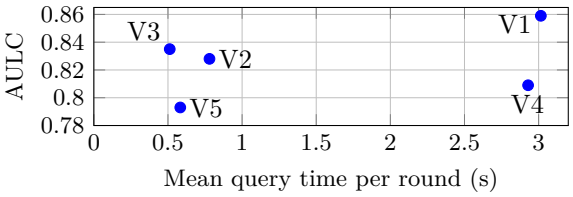


Figure 7. AULC over round vs mean query time per round on the CIFAR100 dataset.

5. Conclusion

We proposed **Conformal Cross-Modal Acquisition (CCMA)**, an active learning framework that combines conformal teacher–student uncertainty with diversity-aware selection for efficient dataset curation. By leveraging pretrained VLM guidance and calibrated prediction sets, CCMA provides more reliable selection signals than vision-only baselines, yielding strong label efficiency across diverse benchmarks. Our findings also reveal when multimodal guidance is beneficial versus when coverage dominates, offering actionable insight into deploying VLMs for real-world AL.

Future work. We aim to extend CCMA to emerging-class and cross-domain settings, and to explore adaptive teacher–student co-learning for more scalable multimodal active learning.

References

[1] Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations*

and Trends® in Machine Learning, 16(4):494–591, 2023. 1

[2] Jordan Ash and Ryan P Adams. On warm-starting neural network training. *Advances in neural information processing systems (NeurIPS)*, 33:3884–3894, 2020. 12

[3] Wonho Bae, Gabriel L. Oliveira, and Danica J. Sutherland. Uncertainty herding: One active learning method for all label budgets. In *International Conference on Learning Representations (ICLR)*, 2025. 2

[4] Jihwan Bang, Sumyeong Ahn, and Jae-Gil Lee. Active prompt learning in vision language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27004–27014, 2024. 1, 3

[5] Javad Zolfaghari Bengar, Joost van de Weijer, Bartlomiej Twardowski, and Bogdan Raducanu. Reducing label effort: Self-supervised meets active learning. In *IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 1631–1639, 2021. 1, 3

[6] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, 2024. 1, 3

[7] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV)*, pages 446–461, 2014. 5, 7, 19

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1877–1901, 2020. 1

[9] Liangyu Chen, Yutong Bai, Siyu Huang, Yongyi Lu, Bihan Wen, Alan Yuille, and Zongwei Zhou. Making your first choice: to address cold start problem in medical active learning. In *Medical Imaging with Deep Learning*, pages 496–525, 2024. 1

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 5

[11] Talfan Evans, Shreya Pathak, Hamza Merzic, Jonathan Schwarz, Ryutaro Tanno, and Olivier J Henaff. Bad students make great teachers: Active learning accelerates large-scale visual understanding. In *European Conference on Computer Vision (ECCV)*, pages 264–280, 2024. 1

[12] Reza Zanjirani Farahani and Masoud Hekmatfar. *Facility location: Concepts, Models, algorithms and case studies*. Springer Science & Business Media, 2009. 2

[13] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006. 5

[14] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In *International Conference on Machine Learning (ICML)*, pages 1183–1192, 2017. 2

[15] Gregory Griffin, Alex Holub, Pietro Perona, et al. Caltech-256 object category dataset. Technical report, California Institute of Technology Pasadena, 2007. 5

[16] Sanket Rajan Gupte, Josiah Aklilu, Jeffrey J Nirschl, and Serena Yeung-Levy. Revisiting active learning in the era of vision foundation models. *Transactions on Machine Learning Research (TMLR)*, 2024. 1, 3, 5

[17] Yeho Gwon, Sehyun Hwang, Hoyoung Kim, Jungseul Ok, and Suha Kwak. Enhancing cost efficiency in active learning with candidate set query. *Transactions on Machine Learning Research (TMLR)*, 2025. 3

[18] Guy Hacohen, Avihu Dekel, and Daphna Weinshall. Active learning on a budget: Opposite strategies suit high and low budgets. In *International Conference on Machine Learning (ICML)*, pages 8175–8195, 2022. 2, 15

[19] Neil Houlsby, Ferenc Huszar, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *Computing Research Repository (CoRR)*, 2011. 2

[20] Cédric Jung, Shirin Salehi, and Anke Schmeink. Active learning with alternating acquisition functions: Balancing the exploration-exploitation dilemma. In *2024 IEEE International Conference on Big Data (BigData)*, pages 5755–5764, 2024. 2

[21] Vishal Kaushal, Rishabh Iyer, Suraj Kothawade, Rohan Mahadev, Khoshrav Doctor, and Ganesh Ramakrishnan. Learning from less data: A unified data subset selection and active learning framework for computer vision. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1289–1299, 2019. 1

[22] Zahra Kharazian, Tony Lindgren, Sindri Magnusson, and Henrik Boström. Copal: Conformal prediction in active learning an algorithm for enhancing remaining useful life estimation in predictive maintenance. In *Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications*, pages 195–217, 2024. 1, 3

[23] Andreas Kirsch, Sebastian Farquhar, Parmida Atighehchian, Andrew Jesson, Frédéric Branchaud-Charron,

and Yarin Gal. Stochastic batch acquisition: A simple baseline for deep active learning. *Transactions on Machine Learning Research (TMLR)*, 2023. 2

[24] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. 5, 7, 19

[25] David D. Lewis and Jason Catlett. Heterogenous uncertainty sampling for supervised learning. In *International Conference on International Conference on Machine Learning (ICML)*, page 148–156, 1994. 2

[26] Xiongquan Li, Xukang Wang, Xuhesheng Chen, Yao Lu, Hongpeng Fu, and Ying Cheng Wu. Unlabeled data selection for active learning in image classification. *Scientific Reports*, 14(1):424, 2024. 1

[27] Lázaro Emílio Makili, Jesús A. Vega Sánchez, and Sebastián Dormido-Canto. Active learning using conformal predictors: Application to image classification. *Fusion Science and Technology*, 62(2):347–355, 2012. 3

[28] Sergio Matiz and Kenneth E. Barner. Conformal prediction based active learning by linear regression optimization. *Neurocomputing*, 388:157–169, 2020. 1, 2, 3

[29] Sayak Nag, Udita Ghosh, Calvin-Khang Ta, Sarosij Bose, Jiachen Li, and Amit K Roy-Chowdhury. Conformal prediction and mllm aided uncertainty quantification in scene graph generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11676–11686, 2025. 1

[30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024. 1, 5, 12, 13

[31] Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Gholamreza (Reza) Haffari, Anton van den Hengel, and Javen Qinfeng Shi. Active learning by feature mixing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12237–12246, 2022. 2

[32] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1406–1415, 2019. 5, 7, 19

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 1, 2, 5, 12

[34] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning (ICML)*, pages 8821–8831, 2021. 1

[35] Oscar Reyes, Abdulrahman H Altalhi, and Sebastián Ventura. Statistical comparisons of active learning strategies over multiple datasets. *Knowledge-Based Systems*, 145:274–288, 2018. 8

[36] Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *International Conference on Machine Learning (ICML)*, pages 441–448, 2001. 2

[37] Bardia Safaei and Vishal M Patel. Active learning for vision-language models. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4902–4912, 2025. 3, 4

[38] Shirin Salehi and Anke Schmeink. Is active learning green? an empirical study. In *IEEE International Conference on Big Data (BigData)*, pages 3823–3829, 2023. 1

[39] Shirin Salehi and Anke Schmeink. Data-centric green artificial intelligence: A survey. *IEEE Transactions on Artificial Intelligence*, 5:1973–1989, 2023. 1

[40] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[41] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. 2

[42] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research (JMLR)*, 9(3), 2008. 1, 2

[43] Julio Silva-Rodríguez, Ismail Ben Ayed, and Jose Dolz. Conformal prediction for zero-shot models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19931–19941, 2025. 1

[44] Alex Tamkin, Dat Nguyen, Salil Deshpande, Jesse Mu, and Noah Goodman. Active learning helps pretrained models learn the intended task. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:28140–28153, 2022. 1

[45] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer US, 2005. 1, 2

[46] Dan Wang and Yi Shang. A new active labeling method for deep learning. *International Joint Conference on Neural Networks (IJCNN)*, pages 112–119, 2014. 2

[47] Jiangyi Wang and Na Zhao. Uncertainty meets diversity: A comprehensive active learning framework for indoor 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20329–20339, 2025. 2

[48] Ofer Yehuda, Avihu Dekel, Guy Hacohen, and Daphna Weinshall. Active learning through a covering lens.

Advances in Neural Information Processing Systems (NeurIPS), 35:22354–22367, 2022. 2, 15

[49] Tianxiang Yin, Ningzhong Liu, and Han Sun. Towards cost-effective learning: A synergy of semi-supervised and active learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10163–10172, 2025. 3