EMBRACING BIASED TRANSITION MATRICES FOR COMPLEMENTARY-LABEL LEARNING WITH MANY CLASSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Complementary-label learning (CLL) is a weakly supervised paradigm for multiclass classification, where each training instance comes with a complementary label indicating a class that the instance does *not* belong to. After nearly a decade of research, CLL still faces a long-standing scalability barrier: as the number of classes grows from the scale of CIFAR-10 to that of TinyImageNet-200, existing methods learn almost nothing meaningful. We show that this barrier is not intrinsic to the problem, but tied to a common modeling choice. Most approaches assume a uniform transition matrix when modeling complementary-label generation, while the potential of assuming or leveraging a biased (non-uniform) transition matrix remains underexplored, despite some empirical successes on real-world data. Through a systematic study, we demonstrate that performance can be significantly improved by deliberately designing a biased transition matrix that restricts complementary labels to a subset of possible classes. Motivated by this finding, we propose *bias-induced constrained labeling* (BICL), a principled and practical framework spanning from data collection to training that properly leverages such bias for solving CLL. Extensive experiments demonstrate that BICL substantially improves CLL performance and, most importantly, breaks this barrier, enabling effective learning on CIFAR-100 and TinyImageNet-200 with more than sevenfold accuracy improvements.

1 INTRODUCTION

Large and accurately annotated data have driven the development of machine learning in recent decades. To address the high cost and limited efficiency of human labeling, weakly supervised learning (WSL) leverages *imperfect supervision*, i.e., training signals that are incomplete, ambiguous, or noisy, yet cheaper to acquire than ground-truth labels via humans or agents (Mai et al., 2025). WSL contains various scenarios (Sugiyama et al., 2022), including positive-unlabeled learning (Elkan & Noto, 2008; Kiryo et al., 2017), confidence information (Cao et al., 2021; Wang et al., 2023), pairwise relationship (Bao et al., 2018; 2022), complementary-label learning (CLL) (Ishida et al., 2017; Chou et al., 2020), partial-label learning (Jin & Ghahramani, 2002; Lv et al., 2020), or noisy-label learning data (Natarajan et al., 2013; Patrini et al., 2017).

In this work, we focus on CLL, where weak supervision specifies a class that an instance does *not* belong to. CLL has attracted increasing attention (Mai et al., 2025; Ishida et al., 2017; Wang et al., 2024b) as it is potentially appealing when identifying an *incorrect* or *opposite* label is easier than determining the true class, making it promising for real-world applications such as crowdsourcing and preference-based feedback (Wang et al., 2025). For example, in crowdsourcing, annotators may find it easier to label a tomato as *not an animal* than to decide whether it should be categorized as a "vegetable" or a "fruit". However, CLL remains one of the most challenging WSL paradigms because each label provides little and indirect negative evidence about the true class. As the label space grows, this evidence becomes increasingly ambiguous, and many existing methods struggle to scale beyond small or moderate settings, such as 10 or 20 classes (Mai et al., 2025; Ishida et al., 2017; Chou et al., 2020; Wang et al., 2025), hindering practical deployment.

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

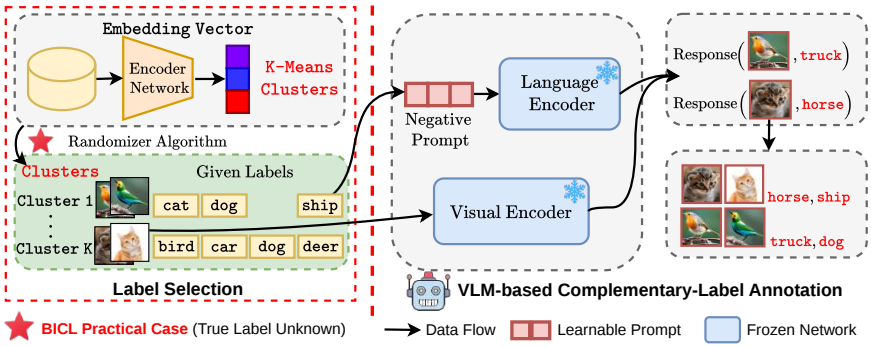


Figure 1: **BICL Practical Case**: Overview of the proposed practical design for bias-induced constrained labeling (BICL) that operates without true label access.

Existing CLL models generally assume that complementary labels (CLs) are generated from the true class according to a transition matrix (Ishida et al., 2017; Yu et al., 2018; Kim et al., 2019; Gao & Zhang, 2021; Lin & Lin, 2023), which plays a central role in CLL studies. Ishida et al. (2017) pioneered the theoretical study of CLL by assuming a zero-diagonal uniform transition matrix, under which each label complementary to the true class is generated with equal probability. Yu et al. (2018) relaxed the assumption by introducing a biased (non-uniform) transition matrix to capture systematic biases during annotation. Such biased transition matrices better reflect realistic annotation mechanisms and have often yielded stronger empirical performance than the uniform ones (Ishida et al., 2017), as observed in subsequent studies (Yu et al., 2018; Ye et al., 2024).

Despite these empirical gains, biased transition matrices remain hard to exploit in large-scale CLL settings such as 100-class classification. The reason is that it remains unclear how to effectively control the bias during data collection. Existing CLL pipelines offer no systematic way to *design* annotation protocols that induce a helpful bias in the transition matrix. In this work, we take a first step with a systematic study and reveal the potential of an oracle annotation process that concentrates on a few candidate CLs per true class. We then move from the oracle process to a *practical* one by proposing *Bias-Induced Constrained Labeling* (BICL), a principled framework that introduces and controls a helpful bias during CLL data collection. As illustrated on the left-hand side of Figure 1, BICL concentrates on a few candidate CLs per *cluster* instead of per true class, allowing it to approximate the oracle design without access to the true labels.

With the practical BICL framework, CLL coupled with few-shot true labels, which is the standard experiment setting in CLL studies, is now on par with ordinary supervised learning that uses many more true labels in many-class settings. In particular, on TinyImageNet-200, BICL-collected complementary labels plus 5 true labels per class (used to estimate the transition matrix (Yu et al., 2018; Lin & Lin, 2023)) can match fully supervised learning with more than 100 true labels per class (Appendix D.8). To our knowledge, this is the first time that complementary labels, which are much weaker, can replace much of the more expensive true labels when building a classifier. The result suggests a new paradigm of building classifiers from *very few strong labels plus many cheap weak labels*.

Our main contributions are summarized as follows.

- We demonstrate that BICL addresses a *long-standing scalability bottleneck* in nearly a decade of CLL research. While prior CLL studies have largely focused on problems with *at most 20* classes, BICL achieves a $7\times$ improvement on CIFAR-100 and an $8\times$ improvement on TinyImageNet-200 compared to Ishida et al. (2017).
- We characterize complementary-label supervision information-theoretically: uniform labeling is maximally ambiguous under balanced classes and carries vanishing information as classes grow, whereas suitably structured transitions retain substantially more.
- Through extensive ablations, we show that BICL induces substantially more informative transition structures, that the gain stems from the cluster-shared candidate sets rather than from the per-image choice of the VLM, and that transition-aware CLL algorithms can exploit these structures effectively in many-class settings.

2 BACKGROUND

2.1 ORDINARY-LABEL CLASSIFICATION

In the ordinary classification setting, we assume that a d-dimensional pattern $\mathbf{x} \in \mathbb{R}^d$ and its corresponding class label $y \in \mathcal{Y} = \{1, \dots, C\}$ are sampled jointly from an unknown underlying density $p(\mathbf{x}, y)$. The dataset is denoted as $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where n represents the number of training data, and these n instance-label pairs are sampled independently from $p(\mathbf{x}, y)$.

The goal of ordinary multi-class classification is to learn a classifier that minimizes the classification risk. To achieve this, we employ a prediction function vector $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \dots, g_C(\mathbf{x}))^\top$, where $^\top$ denotes the transpose and $g_k(\mathbf{x})$ acts as the prediction score for class k. The final classifier $f(\mathbf{x})$ is typically determined by

$$f(\mathbf{x}) = \underset{k \in \{1,\dots,C\}}{\arg\max}\ g_k(\mathbf{x}).$$

The classification risk is defined as the expectation of a multi-class loss function $\mathcal{L}(\mathbf{g}(\mathbf{x}), y)$ computed on the soft label vector $\mathbf{g}(\mathbf{x})$ over the joint density $p(\mathbf{x}, y)$:

$$R(\mathbf{g}) = \mathbb{E}_{p(\mathbf{x},y)}[\mathcal{L}(\mathbf{g}(\mathbf{x}), y)],$$

where $\mathbb{E}$ denotes the expectation. In the context of Empirical Risk Minimization (ERM), since $p(\mathbf{x}, y)$ is unknown, we approximate the risk using the training sample average.

2.2 COMPLEMENTARY-LABEL CLASSIFICATION

In the setting of complementary-label classification, instead of a ground-truth label y, we observe a complementary label $\bar{y}$, which specifies a class that the pattern $\mathbf{x}$ does *not* belong to. Formally, given a complementary-label training set $\bar{\mathcal{D}} = \{(\mathbf{x}_i, \bar{y}_i)\}_{i=1}^{n_{\text{tr}}}$, our goal is to learn a classifier f such that the classification error on an unseen ordinary-label test set $\mathcal{D}_{\text{te}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_{\text{te}}}$ is minimized.

To utilize such data effectively, we often model the relationship between the latent true labels and the observed CLs using a transition matrix $Q \in [0,1]^{C \times C}$, where $Q_{kj} = p(\bar{Y} = j \mid Y = k)$. Since a complementary label cannot be the true label, we typically assume $Q_{kk} = 0$. This transition model allows us to map the ordinary class posteriors $\boldsymbol{\eta}(\mathbf{x}) = (p(Y = 1|\mathbf{x}), \dots, p(Y = C|\mathbf{x}))^\top$ to the complementary posteriors $\bar{\boldsymbol{\eta}}(\mathbf{x}) = (p(\bar{Y} = 1|\mathbf{x}), \dots, p(\bar{Y} = C|\mathbf{x}))^\top$ via $\bar{\boldsymbol{\eta}}(\mathbf{x}) = Q^\top \boldsymbol{\eta}(\mathbf{x})$.

Transition-matrix-based methods, such as forward correction (FWD) (Yu et al., 2018), exploit this relationship by applying the transition operator to model predictions before computing the loss on CLs. In the biased-CL setting, these methods typically assume either a known transition matrix or an estimate obtained from a small clean seed set of ordinarily labeled examples. Existing CLL protocols commonly use only 2 to 5 ordinary examples per class for this purpose (Yu et al., 2018; Lin & Lin, 2023) (Appendix D.7), making limited clean supervision a practical assumption for transition-matrix estimation. Given such an estimate, FWD can exploit the biased transition structure encoded in Q rather than assuming uniform complementary-label generation, thereby improving learning when the observed CLs follow a systematic non-uniform mechanism.

An important distinction lies in how different CLL methods exploit the transition matrix. URE-based methods (Ishida et al., 2017) recover the ordinary classification risk through inverse correction involving Q^{-1}; when the estimated Q is strongly biased or poorly conditioned, matrix inversion can amplify estimation errors and increase the variance of the resulting risk estimates. In contrast, FWD and transition-aware CPE variants exploit Q through a forward mapping.

Alternatively, Ishida et al. (2019) proposed a direct risk minimization framework without estimating the transition matrix under a stronger assumption that the complementary labels are generated *uniformly*. Specifically, this setting assumes that CLs are sampled uniformly from the set of incorrect classes, meaning $Q_{kj} = \frac{1}{C-1}$ for all $j \neq k$. Under this assumption, they derived an unbiased estimator for the classification risk under a moderate condition:

$$R(\mathbf{g}) = (C-1)\mathbb{E}_{p(\mathbf{x},\bar{y})}[\bar{\mathcal{L}}(\mathbf{g}(\mathbf{x}), \bar{y})] + M,$$

where $\bar{\mathcal{L}}$ is the complementary loss computed on $\mathbf{g}(\mathbf{x})$ and M represents a constant. This formulation allows for consistent risk minimization using only CL data.

3 BICL: BIAS-INDUCED CONSTRAINED LABELING

In this section, we present BICL, a structured biased complementary-label collection protocol with two variants: the *BICL Analysis Case* (BICL-A) and the *BICL Practical Case*. To the best of our knowledge, this is the first CLL data-collection study at the 200-class scale, whereas prior work has focused on datasets with at most 20 classes (Mai et al., 2025; Wang et al., 2025; Ye et al., 2024).

3.1 BIASED COMPLEMENTARY-LABEL COLLECTION PROTOCOL

Applying CLL to large-scale classification remains a major bottleneck, raising the question: *Is CLL practical in many-class settings*? Existing studies focus on datasets with at most 20 classes (Mai et al., 2025; Wang et al., 2025), as performance on larger benchmarks is poor. For example, on CIFAR-100, CLL methods typically achieve only 5 to 7 percentage points of accuracy, even when CLs are synthetically generated under uniform and noiseless assumptions (Ishida et al., 2017).

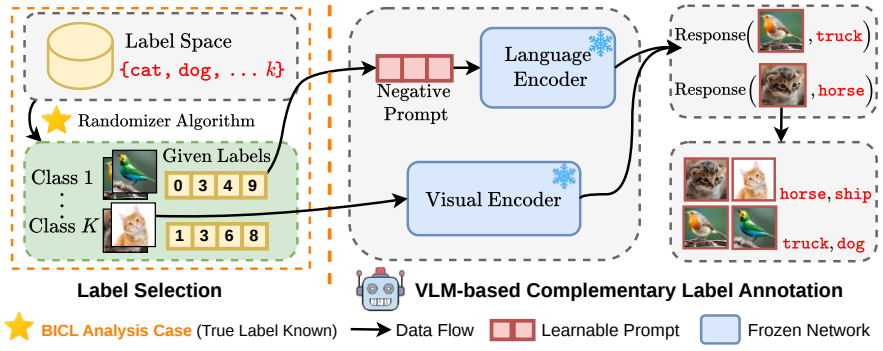


Figure 2: **BICL Analysis Case:** Overview of the proof-of-concept design for complementary candidate-label selection that operates with true label access. It consists of (1) reducing candidate label selection, (2) VLM-based complementary-label annotation with a *negative prompt*.

Extending CLL to many-class settings raises a key question: "*what is the main difficulty?*" The challenge is twofold. First, from the labeling perspective, whether using human annotators or vision-language model (VLM) labelers, it is impractical to present hundreds of candidate classes for complementary labeling. Second, even though existing protocols, such as those in CLImage (Wang et al., 2025) and ACLImage (Mai et al., 2025), address the first challenge by restricting the candidate space by randomly selecting a small subset (e.g., four) of CLs from the full C classes, the candidate space is offered on a per-instance basis. Then, the approach still produces a diffuse, highly scattered transition matrix, weakening the learning signal.

Our key idea is to examine whether further constraining the effective choice space can induce a more informative transition structure in CLs. Motivated by this insight, we propose a novel data collection protocol that systematically introduces structured bias into CLs. The overall pipeline, illustrated in Figure 2, comprises two main components, described below.

Label Selection. This component is central to our data-collection protocol and is crucial for effectively collecting biased CLs. Its role is to determine the candidate complementary label set presented to the VLM for annotation, as illustrated on the *left-hand side* of Figure 2. First, for each class, we randomly sample four labels from the $C-1$ possible CLs[1] using a randomization algorithm. The further details of why we sample four labels can be found in Appendix C.5. Second, we assign the same candidate label set (e.g., $\{0, 3, 4, 9\}$) to all images belonging to that class, forming an initial pool of plausible complementary-label candidates. The key idea is to fix a small candidate set per class, thereby restricting the effective label space and inducing a concentrated, class-dependent transition structure in the resulting CLs. Section 4 gives an information-theoretic account of why such structure carries more information about the true class. To prevent the VLM from developing class-specific label preferences, we randomly shuffle the order of candidate labels for each image when requesting annotation from the VLM.

[1]This is a proof-of-concept setting in which the true label is assumed to be excluded from the candidate complementary label set.

VLM-based Complementary-Label Annotation. Next, we employ a VLM (Liu et al., 2023a) as the annotator in our framework. Prior work by Mai et al. (2025) shows that VLM-based annotation can be more efficient and scalable than human annotation for collecting CLs. We therefore adopt a VLM as our starting point, while noting that the same protocol can be naturally extended to human annotators. Following Mai et al. (2025), we adopt the *negative prompt*: "***Question:***<*image*> *Which label does not belong to this image? Answer the question with a single word: {label[0], label[1], label[2], label[3]}*". For each image, the VLM is prompted to select a label $\bar{y}$ that *does not* correspond to the image content from the candidate set determined by the *label selection* component. This procedure produces a *biased complementary-label dataset*, consisting of image–label pairs where the assigned label is a complementary label (i.e., an image of a bird labeled as a "truck"), while the overall annotation distribution is systematically shaped by our structured candidate-label design. Additional analyses on the effects of VLM choice and negative-prompt variants are provided in Appendices D.4 and D.5, respectively.

To systematically evaluate the effectiveness of *BICL Analysis Case* approach illustrated in Figure 2, we conduct an ablation study by benchmarking three complementary-label generation assumptions: Uniform (Ishida et al., 2017), Biased (Yu et al., 2018), and BICL (ours). Experiments are performed on CIFAR-10 and CIFAR-20 using two representative CLL algorithms, complementary probability estimation (CPE) (Lin & Lin, 2023) and forward correction (FWD) (Yu et al., 2018). The empirical results reported in Table 1 demonstrate that our BICL-based approach consistently outperforms both Uniform and Biased settings across datasets and algorithms. In particular, BICL achieves improvements of more than 6 percentage points and 3 percentage points over the Biased setting on CIFAR-10 and CIFAR-20, respectively. The performance gap is even larger when compared to Uniform CLs, exceeding 19 percentage points on CIFAR-10 and 50 percentage points on CIFAR-20.

Table 1: Accuracy (mean±std) on CIFAR-10/20 for FWD and CPE-F under Uniform, Biased, and our BICL Analysis Case (BICL-A) complementary-label assumptions. Δ is the absolute gain of BICL over the stronger baseline between Uniform and Biased. Best results are bolded.

Datasets	**CIFAR-10**				**CIFAR-20**			
Algorithm	Uniform	Biased	**BICL-A**	Δ	Uniform	Biased	**BICL-A**	Δ
FWD	66.38±2.16	80.04±1.88	**86.31**±0.05	↑ 6.27	20.50±0.75	67.79±0.36	**71.32**±0.25	↑ 3.53
CPE-F	66.24±2.04	79.59±1.87	**85.71**±0.18	↑ 6.12	20.73±1.17	68.40±0.19	**71.58**±0.21	↑ 3.18

Biased: Following Yu et al. (2018), we consider a biased transition setting. As the original transition matrix is not publicly available, we adopt the reproducible construction of (Gao & Zhang, 2021), assigning three probability levels ($\frac{0.75}{3}$, $\frac{0.24}{3}$, $\frac{0.01}{3}$) to CLs for each class y as also used in (Ye et al., 2024).

These results validate the effectiveness of the oracle-based formulation shown in Figure 2. However, it is important to emphasize that this setting is *not deployable* in practice, as it assumes access to the true label to construct noise-free CLs.[2]. In realistic scenarios, such oracle access is unavailable, and annotation noise is unavoidable.

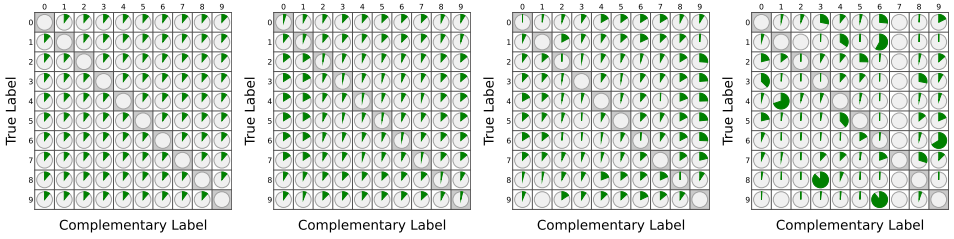
3.2 Practical BICL Design

To move toward a practical and deployable solution, we extend the oracle formulation in Figure 2 to the realistic protocol in Figure 1, which we refer to as the *BICL Practical Case* (hereafter, we use *BICL* for brevity). This practical variant is designed to operate without access to true labels. The key distinction from the idealized framework lies in the ***label selection*** strategy. In standard CLL, true labels are not available during training (Ishida et al., 2017). Consequently, unlike the *BICL Analysis Case*, we cannot group samples by their true class to assign a shared candidate set and then sample CLs accordingly. Moreover, prior work shows that CLL is highly sensitive to annotation noise, which can substantially degrade performance due to overfitting (Mai & Lin, 2026). These considerations motivate the need for a practical mechanism that (i) estimates groups of samples likely to share the same underlying class and (ii) better controls the noise introduced during annotation.

We adopt representations from a pre-trained *encoder network* (Chen & He, 2021) to group samples with similar feature characteristics. This clustering step assigns a cluster index to each sample and

[2]All three dataset assumptions in Table 1 rely on oracle knowledge of the true label to exclude it from the complementary-label set, resulting in a zero transition probability for the true class

serves as a preprocessing phase for candidate label selection. We provide additional analysis on how the choice of encoder affects BICL performance in Appendix D.3. As illustrated in Figure 1, we cluster these representations using K-means to obtain groups of visually similar samples. Each image is then assigned to a cluster Z_c, for $c \in \{1, \ldots, C\}$. We provide an in-depth analysis of the optimal number of clusters for each dataset in Appendix C.4. For each cluster Z_c, we randomly sample a fixed-size candidate label set (four labels in our experiments) from the global label space. When querying the VLM for complementary-label annotation, we randomly shuffle the order of candidate labels for each image to prevent order-induced preferences.



(a) CIFAR-10 (b) CLCIFAR-10 (c) ACLCIFAR-10 (d) BCLCIFAR-10

Figure 3: BICL transition matrix compared to other label collection approaches.

Concentrated transition structure. Figure 3 compares the empirical transition matrix of our CIFAR-10 dataset (BCLCIFAR-10) with the idealized uniform matrix (Ishida et al., 2017), the human-annotated CLCIFAR-10 (Wang et al., 2025), and the VLM-annotated uniform design ACLCIFAR-10 (Mai et al., 2025): ours is far more concentrated, as a direct consequence of constrained labeling combined with the systematic preferences of the VLM. The ablation in Appendix C.2 shows that the gains come primarily from the constrained-labeling design rather than from VLM preferences alone, and Appendix C.7 reports label-noise rates, further transition matrices, and label distributions for all BICL datasets.

4 An Information-Theoretic View of Complementary Labels

A complementary label says one class an image is not; what that is worth depends on how the label was produced. An annotator who names any wrong class at random rules out one class in C, whereas one who answers from a small, class-dependent candidate set narrows the true class down much more. We make this precise by treating the labeling mechanism as a channel from the true class Y to $\bar{Y}$ and asking how much information about Y gets through. The analysis characterizes the supervision signal, not the error of any particular learner; proofs are in Appendix A.

Setup. With the transition matrix $Q_{ij} = p(\bar{Y} = j \mid Y = i)$ of Section 2.2 and the class prior $p_i = p(Y = i)$, the joint distribution is $p_i Q_{ij}$. Ambiguity is $H(Y \mid \bar{Y})$, the uncertainty about the true class left by one complementary label, and information is $I(Y; \bar{Y}) = H(Y) - H(Y \mid \bar{Y})$, both in bits. The prior is *balanced* when $p_i = 1/C$, as in our benchmarks and the training labels of Section 5; Q is *valid* when $Q_{ii} = 0$ and every row sums to one, which the collected matrices satisfy up to a diagonal noise mass of 0.03–0.8%. Throughout we use

$$H(Y \mid \bar{Y}) \;=\; H(Y) + H(\bar{Y} \mid Y) - H(\bar{Y}), \tag{1}$$

which separates the within-class output uncertainty $H(\bar{Y} \mid Y)$ from the global output diversity $H(\bar{Y})$: a complementary label is informative when each class emits a concentrated set of labels while the mechanism as a whole keeps a diverse set of outputs, two requirements that can pull in opposite directions.

Theorem 1 (Maximum ambiguity of uniform complementary labeling)**.** *Assume a balanced prior and a valid Q. Then $H_Q(Y \mid \bar{Y}) \le \log(C-1)$, with equality if and only if $Q = Q^{\mathrm{unif}}$, where $Q^{\mathrm{unif}}_{ij} = \frac{1}{C-1}$ for $i \neq j$ (Ishida et al., 2017). Hence $H_{\mathrm{unif}}(Y \mid \bar{Y}) = \log(C-1)$ and $I_Q(Y; \bar{Y}) \ge I_{\mathrm{unif}}(Y; \bar{Y}) = \log \frac{C}{C-1}$, with equality iff $Q = Q^{\mathrm{unif}}$.*

Uniform labeling is the most ambiguous. A uniform complementary label j excludes only class j and spreads the remaining posterior mass evenly over the other $C - 1$ classes; any other valid

Table 2: Supervision ambiguity and information per complementary label (bits, balanced prior) for uniform labeling and the collected BICL matrices in the analysis (BICL-A) and practical (BICL) cases. Cert.: corollary certifying lower ambiguity than uniform labeling (1: Cor. 1; 2: Cor. 2, $k_0 = 4$), with the largest row support k in parentheses.

		$H(Y \mid \bar{Y})$			$I(Y;\bar{Y})$			Cert. (k)	
Dataset	C	Unif.	BICL-A	BICL	Unif.	BICL-A	BICL	BICL-A	BICL
CIFAR-10	10	3.17	**1.80**	2.20	0.152	**1.52**	1.12	1,2 (4)	2 (10)
CIFAR-20	20	4.25	**1.78**	3.62	0.074	**2.54**	0.70	1,2 (5)	2 (19)
CIFAR-100	100	6.63	**1.83**	4.57	0.014	**4.82**	2.07	1,2 (4)	1,2 (68)
TinyImageNet-200	200	7.64	**1.89**	5.84	0.007	**5.75**	1.80	1,2 (4)	2 (78)

mechanism concentrates that mass somewhere (Appendix A.1, which also shows that the balanced prior is needed only for the optimality claim, not for the bound). Since $\log \frac{C}{C-1} \approx \frac{1}{C \ln 2}$, the information in one uniform complementary label decays like $1/C$, from 0.152 bits at $C = 10$ to 0.007 bits at $C = 200$: at the scale of CIFAR-100 and TinyImageNet-200, uniform CLL has little to work with at the level of the mechanism, however capable the learner. The gain of a non-uniform mechanism depends on the structure of Q, which the next result makes explicit.

Theorem 2 (Information bound for k-sparse complementary-label mechanisms)**.** *If every row of Q has support size at most k, then $H(\bar{Y} \mid Y) \le \log k$; hence $I(Y;\bar{Y}) \ge H(\bar{Y}) - \log k$ and $H(Y \mid \bar{Y}) \le H(Y) + \log k - H(\bar{Y})$. Under a balanced prior, with the* diversity deficit $\delta := \log C - H(\bar{Y}) = D_{\mathrm{KL}}(p(\bar{Y}) \,\|\, \mathrm{Unif}) \ge 0$, $H(Y \mid \bar{Y}) \le \log k + \delta$.

Sparse rows help only if the outputs stay diverse. Theorem 2 splits ambiguity into two competing terms. Restricting each class to k candidates caps the within-class term $H(\bar{Y} \mid Y)$ at $\log k$, which is why constraining the candidate set can help; the diversity deficit δ is the price paid when the labels of different classes pile up on the same few columns. Both must be small at once: tight rows that all point at the same handful of labels have a small k but a large δ, and naive sparsification can even raise ambiguity, by 1.9 bits in the 10-class example of Appendix A.4. BICL is therefore designed to induce not sparsity as such but concentrated class-conditional outputs whose supports differ across classes. Corollary 1 turns this into a margin over uniform labeling whenever $\delta < \log \frac{C-1}{k}$, and Corollary 2 relaxes the exact support condition to a small mass outside the k_0 largest entries of each row, the form needed after clustering (Appendix A.3); neither uses the diagonal of Q, so both apply to the collected matrices as measured.

BICL through this lens. BICL assigns four candidate labels, drawn at random from the full label space, to every class (analysis case) or cluster (practical case), concentrating the rows of Q while steering different classes toward different columns. Table 2 shows the effect on the collected matrices. In the analysis case, ambiguity stays at about 1.8 bits from 10 to 200 classes while uniform ambiguity grows as $\log(C-1)$: a small candidate set keeps the information in one label from vanishing with C, because the number of classes that can produce a given label is governed by k rather than C (Appendix A.5). The practical case weakens this structure, since a cluster mixes several true classes and hands all of them the same candidate set, so supports broaden and ambiguity rises relative to the oracle design. Nevertheless, the relaxed bound certifies lower ambiguity than uniform labeling on all four benchmarks, and the information per label on TinyImageNet-200 rises from 0.007 to 1.80 bits (the factor of roughly 7 to 250 across datasets is large mainly because the uniform value approaches zero). This quantifies the price of replacing true-class grouping with deployable feature-based clustering; Section 5 evaluates whether learning algorithms can exploit the resulting structure.

5 Experiments

In this section, we present the experimental setup and compare BICL with representative baselines to evaluate its effectiveness, robustness, and scalability.

5.1 Experimental Setups

Datasets and implementation details. We evaluate on four synthetically labeled datasets, CIFAR-10, CIFAR-20, CIFAR-100 (Krizhevsky, 2009) and TinyImageNet-200 (Le & Yang, 2015), with 10, 20, 100 and 200 classes; each CIFAR dataset has 50,000 training and 10,000 test images,

Table 3: Accuracy (%) under four annotation settings on CIFAR-10 and CIFAR-20; Δ: our **BICL** minus the second-best setting.

Datasets	**CIFAR-10**					**CIFAR-20**				
Algorithm	Uniform	CLImage	ACLImage	**BICL**	Δ	Uniform	CLImage	ACLImage	**BICL**	Δ
FWD	64.99±0.64	39.25±1.55	56.58±1.18	**81.23**±0.05	↑ 16.24	20.50±0.75	19.82±0.38	25.35±0.35	**50.84**±0.32	↑ 25.49
CPE-F	64.97±0.61	38.94±1.39	56.43±0.84	**80.98**±0.18	↑ 16.01	20.73±1.17	19.48±0.35	25.02±0.68	**50.89**±0.30	↑ 25.87
CPE-I	54.44±1.95	33.94±1.40	52.79±1.34	**75.64**±0.59	↑ 21.20	12.01±0.93	16.79±0.48	21.58±0.97	**46.42**±0.54	↑ 24.84
CPE-T	57.94±0.97	38.80±1.16	52.84±1.05	**77.47**±1.28	↑ 19.53	20.58±0.78	19.33±0.75	23.39±0.68	**45.31**±0.66	↑ 21.92
URE-TNN	**47.25**±2.35	31.21±1.11	30.70±4.56	42.84±2.27	↓ 4.41	15.29±1.11	9.47±2.67	8.63±1.70	**17.33**±1.68	↑ 2.02
URE-TGA	**57.37**±1.12	33.68±1.06	42.28±5.19	45.65±6.72	↓ 11.72	**13.61**±4.51	5.17±0.25	5.03±0.10	5.05±0.00	↓ 8.56
Std-Supp	85.09±0.39					33.19±0.21				

Std-Supp (Standard Supervision): true labels collected with the same VLM instead of CLs and trained with cross-entropy, to compare standard supervision with CLL.

TinyImageNet-200 has 100,000 and 10,000. All experiments use a ResNet34 (He et al., 2016) backbone with *RandomHorizontalFlip*, *RandomCrop* and normalization, trained with Adam (learning rate 10^{-3}, weight decay 10^{-5}), batch size 512 and 300 epochs on an NVIDIA A6000 48GB GPU. The experiments were repeated three times with different random seeds to ensure robustness.

Baseline methods. We consider six representative CLL algorithms: FWD (Yu et al., 2018), CPE-F/I/T (Lin & Lin, 2023), and URE-TNN/TGA (Ishida et al., 2019). FWD and CPE-F/T apply a forward transition mapping, CPE-I does not explicitly use Q, and URE reconstructs the ordinary risk through inverse correction with Q^{-1}. For methods requiring an estimated transition matrix, we use five ground-truth labels per class; details of the estimation procedure are in Appendix D.7. Related work and the complete experimental setup are provided in Appendices B and E.

5.2 Results and Analysis

As shown in Tables 3 and 4, BICL consistently yields substantial improvements across a wide range of CLL algorithms under four annotation settings: the idealized uniform assumption, human-annotated CLCIFAR, VLM-annotated uniform-style ACLCIFAR, and our BCLCIFAR. On CIFAR-10 and CIFAR-20 (Table 3), BICL achieves the best performance for all FWD and CPE-based methods, with gains of approximately 16–21 percentage points on CIFAR-10 and 22–26 on CIFAR-20 over the second-best baselines. The larger gain on CIFAR-20 indicates that BICL is particularly effective in more challenging, finer-grained settings, consistent with Section 4, where BICL matrices carry more information about the true class than uniform labeling, although that analysis does not by itself order test accuracies.

URE-based methods respond unevenly to BICL: URE-TNN benefits in some settings, whereas URE-TGA often performs substantially worse than FWD and CPE-based methods. URE relies on inverse risk correction involving Q^{-1} (Section 2.2), so estimation errors are amplified when the BICL transition matrix is strongly biased or poorly conditioned; URE-TNN partially mitigates this by truncating negative risk estimates, whereas URE-TGA reverses the gradient for negative class-wise risks and is more sensitive to such errors. FWD and CPE-based methods avoid explicit inversion and benefit more consistently, so BICL is particularly effective with algorithms that exploit strongly non-uniform transitions without unstable inverse correction. Importantly, Table 4 shows that BICL enables CLL to scale to label spaces that were previously infeasible: for FWD and CPE-based methods, accuracy rises from near-random under the uniform assumption to 44–47 percentage points on CIFAR-100 and above 30 on TinyImageNet-200. Structured transitions thus matter most when the label space is large, where uniform complementary labels carry almost no information.

Table 4: Accuracy (%) on CIFAR-100 and TinyImageNet-200 under the settings of Table 3; Δ as there.

Datasets	**CIFAR-100**					**TinyImageNet-200**				
Algorithm	Uniform	CLImage	ACLImage	**BICL**	Δ	Uniform	CLImage	ACLImage	**BICL**	Δ
FWD	5.53±0.47	N/A	N/A	**46.70**±0.60	↑ 41.17	4.00±0.43	N/A	N/A	**32.15**±0.30	↑ 28.15
CPE-F	5.06±0.76	N/A	N/A	**46.57**±0.09	↑ 41.51	2.14±0.87	N/A	N/A	**31.89**±0.07	↑ 29.75
CPE-I	5.88±1.27	N/A	N/A	**34.03**±0.15	↑ 28.15	1.67±0.20	N/A	N/A	**21.70**±0.40	↑ 20.03
CPE-T	3.91±1.09	N/A	N/A	**44.35**±0.33	↑ 40.44	1.19±0.06	N/A	N/A	**28.51**±1.40	↑ 27.32
URE-TNN	1.39±0.05	N/A	N/A	**9.58**±1.06	↑ 8.19	0.60±0.05	N/A	N/A	**5.84**±0.49	↑ 5.24
URE-TGA	1.03±0.03	N/A	N/A	**1.05**±0.01	↑ 0.02	**0.54**±0.03	N/A	N/A	0.49±0.00	↓ 0.05
Std-Supp	36.49±0.49					16.54±0.28				

N/A: CLImage and ACLImage do not cover CIFAR-100 or TinyImageNet-200; prior CLL work stops at 20 classes, so no complementary-label version of these benchmarks exists.

Table 5: Where the gain comes from: ambiguity $H(Y \mid \bar{Y})$ (bits) and FWD accuracy (%, mean±std, three training seeds) with the collected clusters and candidate sets kept and only the selector changed.

Selector	$H(Y \mid \bar{Y})$	Acc. CIFAR-20	$H(Y \mid \bar{Y})$	Acc. CIFAR-100
Uniform CL (no candidate sets)	4.25	19.9±0.5	6.63	5.1±0.4
BICL sets, random pick (true label excluded)	3.74	48.5±1.3	4.81	43.7±3.3
BICL sets, LLaVA pick (BICL)	3.62	49.2±0.7	4.57	46.1±2.1

Table 3 also shows a clear discrepancy between the commonly used *Uniform* assumption (Ishida et al., 2017) and CLs collected in practice. Because it assumes noise-free CLs, the uniform transition has been treated as a reasonable default and often as an implicit *upper bound*, and CLImage (Wang et al., 2025) and ACLImage (Mai et al., 2025) collect CLs that approximate it on real-world data; unavoidable annotation noise and systematic collection biases nevertheless leave these "uniform-like" labels substantially below the idealized setting for most algorithms. BICL, in contrast, outperforms CLImage and ACLImage and even exceeds the idealized *Uniform* setting, so carefully controlled bias can yield higher accuracy than uniformly sampled CLs (Tables 3 and 4), and it reaches its peak performance in fewer epochs (Figure 4, Appendix C). These findings challenge the assumption that uniform CLs are the best one can hope for: deliberately structured non-uniform collection can be more effective than uniform collection for scalable CLL.

Where the gain comes from. Relative to uniform labeling, BICL adds cluster-shared candidate sets and a VLM that picks from them; Table 5 keeps the collected clusters and sets and replaces LLaVA by a uniform pick from the set with the true label excluded. This random selector comes within 0.8 and 2.3 points of LLaVA on CIFAR-20 and CIFAR-100 while both exceed uniform labeling by 29 to 41 points, with similar ambiguity, so the gain comes from the cluster-shared candidate sets that shape Q rather than from the per-image choice of the VLM; no learner observes clusters or sets, since each training label is sampled from Q. Naming the clusters with five true labels per class and training with cross-entropy instead gives 66.3% on CIFAR-10 and 30.6% on CIFAR-100, far below BICL in Tables 3–4, so the complementary-label channel adds supervision beyond what the clusters alone provide (further discussion and experiments in Appendices C, D and E).

6 Conclusion

Conclusion. We introduced BICL, a complementary-label collection framework that deliberately shapes the transition structure rather than treating annotation bias only as a nuisance. Our information-theoretic analysis shows that uniform complementary labeling becomes increasingly uninformative as the label space grows, while structured mechanisms retain substantially more information when class-conditional outputs are concentrated without collapsing global label diversity. Empirically, BICL lets transition-aware CLL methods such as FWD and CPE reach 46 percentage points of accuracy on CIFAR-100 and 32 on TinyImageNet-200, up from 6 and 4 under uniform complementary labels, where conventional CLL learns little; the gain traces to the cluster-shared candidate sets that shape Q rather than to the per-image choice of the VLM. The benefit is largest for methods that apply the transition matrix in the forward direction and does not extend to every CLL learner.

Limitations. BICL relies on the encoder and clustering stage; poor clusters make the induced transition structure more diffuse and shrink the downstream advantage (Appendix D.3). Practical BICL also induces an instance-dependent annotation process, whereas the transition-aware learners we use rely on the class-conditional matrix $Q = P(\bar{Y} \mid Y)$, a marginal approximation of that mechanism. Structured transitions do not benefit every CLL algorithm: forward-transition methods such as FWD and CPE gain consistently, whereas inverse-risk methods can become unstable for a strongly non-uniform or poorly conditioned Q (Section 5). Finally, like other Q-aware methods, we estimate Q from a small clean seed set of 2–5 labeled examples per class.

Future directions. Modeling $P(\bar{Y} \mid Y, X)$ directly, CL-based estimation of Q, and objectives robust to a misspecified Q would reduce the dependence on seed labels and on the marginal approximation; human-in-the-loop annotation is the natural next step for understanding how human semantic ambiguity and labeling bias interact with constrained labeling. More broadly, our results suggest that the complementary-label generation mechanism is a design variable rather than a given, and that designing it for the learner is a promising route to scalable CLL.

Ethics Statement

This work studies weakly supervised learning on public image-classification benchmarks (CIFAR-10, CIFAR-20, CIFAR-100 and TinyImageNet-200) and involves no human subjects, personal data, or newly collected images. The complementary labels we release are produced by open-weight vision-language models (VLMs) under the protocol in Section 3.1; we use these models only as annotators of existing benchmark images, and the datasets we build inherit the licenses of their source benchmarks. Because BICL deliberately induces bias in the complementary-label distribution, we note that this bias is a property of the label *collection* protocol, not of any demographic group, and Appendix C.3 examines its effect on minority classes. As with any method that lowers the cost of labeling, BICL could be applied to build classifiers for purposes we do not endorse; we see no risk specific to this work beyond those common to image classification, and we do not foresee any negative societal impact particular to it.

Reproducibility Statement

We take several steps to ensure reproducibility. Full details of the experimental setup, including the VLMs and prompts used for annotation, benchmark datasets, compared approaches, network architectures, and complementary-label learning algorithms with their hyperparameters, are provided in Section 5.1 and Appendix E. Proofs of the theoretical results in Section 4 are provided in Appendix A. The supplementary material includes the source code for the complete BICL pipeline, including label selection, VLM-based annotation, and training, as well as the transition matrices of all collected datasets, configuration files for reproducing every table and figure, and scripts for regenerating figures from the logged results.

AI Use Statement

We used large language models solely to assist with writing and code maintenance. For writing, they were used to check typos, grammar, and phrasing. For code, they were used for minor cleaning, refactoring, and readability improvements, without contributing to the implementation of research methods or experimental design. They were not used to generate research ideas or design experiments; all research contributions were made entirely by the authors, who reviewed the AI-assisted outputs and took full responsibility for the content of the paper and the associated code. Separately, VLMs constitute a component of the proposed method by serving as complementary-label annotators, as described in Section 3.1.

References

Han Bao, Gang Niu, and Masashi Sugiyama. Classification from pairwise similarity and unlabeled data. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 461–470, 2018.

Han Bao, Takuya Shimada, Liyuan Xu, Issei Sato, and Masashi Sugiyama. Pairwise supervision can provably elicit a decision boundary. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pp. 2618–2640, 2022.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, volume 32, pp. 1565–1576, 2019.

Yuzhou Cao, Lei Feng, Yitian Xu, Bo An, Gang Niu, and Masashi Sugiyama. Learning from similarity-confidence data. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 1272–1282, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *IEEE/CVF international conference on computer vision*, pp. 9640–9649, 2021.

Yu-Ting Chou, Gang Niu, Hsuan-Tien Lin, and Masashi Sugiyama. Unbiased risk estimators can mislead: A case study of learning with complementary labels. In *International Conference on Machine Learning*, pp. 1929–1938, 2020.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 113–123, 2019.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 702–703, 2020.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9268–9277, 2019.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 213–220, 2008.

Lei Feng, Takuo Kaneko, Bo Han, Gang Niu, Bo An, and Masashi Sugiyama. Learning with multiple complementary labels. In *International Conference on Machine Learning*, pp. 3072–3081, 2020.

Yi Gao and Min-Ling Zhang. Discriminative complementary-label learning with weighted loss. In *International Conference on Machine Learning*, pp. 3587–3597, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. Learning from complementary labels. In *Advances in Neural Information Processing Systems*, pp. 5639–5649, 2017.

Takashi Ishida, Gang Niu, Aditya Menon, and Masashi Sugiyama. Complementary-label learning for arbitrary losses and models. In *International Conference on Machine Learning*, pp. 2971–2980, 2019.

Hiroki Ishiguro, Takashi Ishida, and Masashi Sugiyama. Learning from noisy complementary labels with robust loss functions. *IEICE Transactions on Information and Systems*, 105:364–376, 2022.

Haoran Jiang, Zhihao Sun, and Yingjie Tian. Comco: Complementary supervised contrastive learning for complementary label learning. *Neural Networks*, 169:44–56, 2024.

Rong Jin and Zoubin Ghahramani. Learning with multiple labels. *Advances in Neural Information Processing Systems*, 15, 2002.

Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. NLNL: Negative learning for noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 101–110, 2019.

Ryuichi Kiryo, Gang Niu, Marthinus C. du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1675–1685, 2017.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Computer Science University of Toronto, Canada, 2009.

Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. Report of CS231N: Deep Learning for Computer Vision Course, 2015. Stanford University.

Wei-I Lin and Hsuan-Tien Lin. Reduction from complementary-label learning to probability estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 469–481, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023a.

Shuqi Liu, Yuzhou Cao, Qiaozhen Zhang, Lei Feng, and Bo An. Consistent complementary-label learning via order-preserving losses. In *International Conference on Artificial Intelligence and Statistics*, pp. 8734–8748, 2023b.

Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. Progressive identification of true labels for partial-label learning. In *International Conference on Machine Learning*, pp. 6500–6510, 2020.

Tan-Ha Mai and Hsuan-Tien Lin. Intra-cluster mixup: An effective data augmentation technique for complementary-label learning. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Tan-Ha Mai, Nai-Xuan Ye, Yu-Wei Kuan, Po-Yi Lu, and Hsuan-Tien Lin. The unexplored potential of vision-language models for generating large-scale complementary-label learning data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 90–102, 2025.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1944–1952, 2017.

Pierre H Richemond, Jean-Bastien Grill, Florent Altché, Corentin Tallec, Florian Strub, Andrew Brock, Samuel Smith, Soham De, Razvan Pascanu, Bilal Piot, et al. Byol works even without batch statistics. *arXiv preprint arXiv:2010.10241*, 2020.

Masashi Sugiyama, Han Bao, Takashi Ishida, Nan Lu, Tomoya Sakai, and Gang Niu. *Machine learning from weak supervision: An empirical risk minimization approach*. MIT Press, 2022.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024. URL https://arxiv.org/abs/2403.08295.

Haobo Wang, Mingxuan Xia, Yixuan Li, Yuren Mao, Lei Feng, Gang Chen, and Junbo Zhao. Solar: Sinkhorn label refinery for imbalanced partial-label learning. *Advances in neural information processing systems*, 35:8104–8117, 2022a.

Haobo Wang, Ruixuan Xiao, Yixuan Li, Lei Feng, Gang Niu, Gang Chen, and Junbo Zhao. PiCO: Contrastive label disambiguation for partial label learning. In *International Conference on Learning Representations*, 2022b.

Hsiu-Hsuan Wang, Mai Tan Ha, Nai-Xuan Ye, Wei-I Lin, and Hsuan-Tien Lin. CLImage: Human-annotated datasets for complementary-label learning. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Wei Wang, Lei Feng, Yuchen Jiang, Gang Niu, Min-Ling Zhang, and Masashi Sugiyama. Binary classification with confidence difference. In *Advances in Neural Information Processing Systems 36*, pp. 5936–5960, 2023.

Wei Wang, Takashi Ishida, Yu-Jie Zhang, Gang Niu, and Masashi Sugiyama. Learning with complementary labels revisited: The selected-completely-at-random setting is more practical. In *International Conference on Machine Learning*, volume 235, pp. 50683–50710, 2024a.

Wei Wang, Takashi Ishida, Yu-Jie Zhang, Gang Niu, and Masashi Sugiyama. Learning with complementary labels revisited: The selected-completely-at-random setting is more practical. In *Proceedings of the 41st International Conference on Machine Learning*, 2024b.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022.

Meng Wei, Yong Zhou, Zhongnian Li, and Xinzheng Xu. Class-imbalanced complementary-label learning via weighted loss. *Neural Networks*, 166:555–565, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Nai-Xuan Ye, Tan-Ha Mai, Hsiu-Hsuan Wang, Wei-I Lin, and Hsuan-Tien Lin. libcll: an extendable python toolkit for complementary-label learning, 2024. URL https://arxiv.org/abs/2411.12276.

Yiwei You, Jinglong Huang, Qiang Tong, and Bo Wang. Tackling biased complementary label learning with large margin. *Information Sciences*, 687:121400, 2025.

Xiyu Yu, Tongliang Liu, Mingming Gong, and Dacheng Tao. Learning with biased complementary labels. In *European Conference on Computer Vision*, pp. 68–83, 2018.

APPENDIX

A PROOFS AND SUPPLEMENTARY RESULTS FOR SECTION 4

This appendix contains the proofs of the results in Section 4, the worked example showing that sparsity by itself is not enough, a short proposition on posterior support that explains why the analysis case is insensitive to the number of classes, and a description of how the numbers in Tables 2 and 7 were computed. Together they support the reading of Section 4: uniform labeling loses information as C grows, structure recovers it only when concentration is paired with diversity, and the collected matrices behave as this picture predicts.

Throughout, $Y \in [C]$ has prior $p_i = p(Y = i)$, $\bar{Y} \in [C]$, $Q_{ij} = p(\bar{Y} = j \mid Y = i)$, $J_{ij} = p_i Q_{ij}$ is the joint, $q_j = \sum_i J_{ij}$ is the marginal of $\bar{Y}$, and $\mathrm{supp}(v) = \{j : v_j > 0\}$. Logarithms are base 2. Two standard facts are used without further comment: a distribution on m points has entropy at most $\log m$, with equality only for the uniform distribution, and $I(Y;\bar{Y}) = H(Y) - H(Y \mid \bar{Y}) = H(\bar{Y}) - H(\bar{Y} \mid Y)$, which is where (1) comes from.

A.1 PROOF OF THEOREM 1

Proof. We show, in turn, that every valid mechanism satisfies $H_Q(Y \mid \bar{Y}) \le \log(C-1)$ (Step 1), that the uniform mechanism attains this value (Step 2), that it is the only mechanism that does so (Step 3), and the consequence for $I_Q(Y;\bar{Y})$ (Step 4).

Step 1. Upper-bound the ambiguity of any valid mechanism. Fix a complementary label j with $q_j > 0$. By Bayes' rule, the posterior of the true class is

$$p(Y = i \mid \bar{Y} = j) = \frac{p_i Q_{ij}}{q_j}, \qquad i \in [C].$$

Since Q is valid, $Q_{jj} = 0$, and therefore

$$p(Y = j \mid \bar{Y} = j) = \frac{p_j Q_{jj}}{q_j} = 0.$$

Hence the posterior is supported on $[C] \setminus \{j\}$, a set of $C-1$ classes. A distribution on at most $C-1$ points has entropy at most $\log(C-1)$ (maximum entropy on a finite support), so

$$H(Y \mid \bar{Y} = j) \le \log(C-1) \qquad \text{for every } j \text{ with } q_j > 0.$$

Labels with $q_j = 0$ receive zero weight in the conditional entropy, so averaging over $\bar{Y}$ gives

$$\begin{aligned} H_Q(Y \mid \bar{Y}) &= \sum_{j:\, q_j > 0} q_j \, H(Y \mid \bar{Y} = j) \\ &\le \sum_{j:\, q_j > 0} q_j \log(C-1) \\ &= \log(C-1). \end{aligned}$$

This step uses only validity; it holds for every prior.

Step 2. Evaluate the uniform mechanism. Let $p_i = 1/C$ and $Q = Q^{\mathrm{unif}}$, that is, $Q_{ij} = \frac{1}{C-1}$ for $i \ne j$ and $Q_{jj} = 0$. The marginal of $\bar{Y}$ is

$$q_j = \sum_{i=1}^{C} p_i Q_{ij} = \frac{1}{C} \sum_{i \ne j} \frac{1}{C-1} = \frac{1}{C} \cdot \frac{C-1}{C-1} = \frac{1}{C} > 0 \qquad \text{for every } j \in [C],$$

so every posterior is defined, and for $i \ne j$

$$p(Y = i \mid \bar{Y} = j) = \frac{p_i Q_{ij}}{q_j} = \frac{\frac{1}{C} \cdot \frac{1}{C-1}}{\frac{1}{C}} = \frac{1}{C-1}.$$

Each posterior is therefore uniform on the $C-1$ classes other than j, so $H(Y \mid \bar{Y} = j) = \log(C-1)$ for every j, and

$$H_{\text{unif}}(Y \mid \bar{Y}) = \sum_{j=1}^{C} q_j \, H(Y \mid \bar{Y} = j) = \sum_{j=1}^{C} \frac{1}{C} \log(C-1) = \log(C-1).$$

Thus the bound of Step 1 is attained by Q^{unif}.

Step 3. Characterize the equality case. Assume $p_i = 1/C$ and $H_Q(Y \mid \bar{Y}) = \log(C-1)$ for a valid Q. Each term of the average in Step 1 satisfies $H(Y \mid \bar{Y} = j) \le \log(C-1)$ and the weights q_j over $\{j : q_j > 0\}$ sum to one, so

$$0 = \log(C-1) - H_Q(Y \mid \bar{Y}) = \sum_{j:\, q_j > 0} q_j \big(\log(C-1) - H(Y \mid \bar{Y} = j) \big)$$

is a sum of nonnegative terms, and every term with $q_j > 0$ must vanish:

$$H(Y \mid \bar{Y} = j) = \log(C-1) \qquad \text{for every } j \text{ with } q_j > 0.$$

By the equality case of the maximum-entropy property, each such posterior is uniform on $[C] \setminus \{j\}$ (an entropy of $\log(C-1)$ on a support of at most $C-1$ points forces the support to be all of $[C] \setminus \{j\}$),

$$p(Y = i \mid \bar{Y} = j) = \frac{1}{C-1} \qquad \text{for all } i \neq j.$$

Under the balanced prior, $p(Y = i \mid \bar{Y} = j) = \frac{(1/C)\, Q_{ij}}{q_j}$, so uniformity of the posterior says that $Q_{ij} = \frac{C q_j}{C-1}$ does not depend on i. Define

$$c_j := \frac{C q_j}{C-1} \ \text{ if } q_j > 0, \qquad c_j := 0 \ \text{ if } q_j = 0.$$

In either case

$$Q_{ij} = c_j \qquad \text{for all } i \neq j,$$

because for $q_j = 0$ the identity $q_j = \frac{1}{C} \sum_i Q_{ij} = 0$ forces every entry of column j to vanish. In words, every column of Q is constant off the diagonal. Now use row stochasticity together with $Q_{ii} = 0$: for every row i,

$$1 = \sum_{j=1}^{C} Q_{ij} = \sum_{j \neq i} Q_{ij} = \sum_{j \neq i} c_j.$$

Writing $S := \sum_{j=1}^{C} c_j$, this reads

$$S - c_i = 1 \qquad \text{for every } i \in [C],$$

hence $c_i = S - 1$ for every i: all the c_i share a common value c. Then $S = Cc$, and the row constraint $S - c = 1$ becomes

$$(C-1)\, c = 1, \qquad \text{i.e.} \qquad c = \frac{1}{C-1} > 0.$$

Therefore $Q_{ij} = \frac{1}{C-1}$ for all $i \neq j$ and $Q_{ii} = 0$, that is, $Q = Q^{\text{unif}}$ (in particular no column has $q_j = 0$). Conversely, Step 2 shows that Q^{unif} attains equality. Hence equality holds if and only if $Q = Q^{\text{unif}}$.

Step 4. Consequence for the mutual information. Under the balanced prior $H(Y) = \log C$, so

$$\begin{aligned} I_Q(Y; \bar{Y}) &= H(Y) - H_Q(Y \mid \bar{Y}) \\ &= \log C - H_Q(Y \mid \bar{Y}) \\ &\ge \log C - \log(C-1) = \log \frac{C}{C-1}, \end{aligned}$$

with equality if and only if $H_Q(Y \mid \bar{Y}) = \log(C-1)$, which by Step 3 means $Q = Q^{\text{unif}}$; and $I_{\text{unif}}(Y; \bar{Y}) = \log C - \log(C-1)$ by Step 2. This completes the proof. □

The role of the two assumptions. The two assumptions play different roles. Validity, $Q_{ii} = 0$ with unit row sums, is what gives the support-based cap $\log(C-1)$, because it confines every posterior to $C-1$ classes; without it the cap fails, since a $\bar{Y}$ that is independent of Y with full support has $H(Y \mid \bar{Y}) = H(Y)$, which equals $\log C$ under a balanced prior. The balanced prior is needed only for the optimality claim, that the uniform mechanism attains the cap and is the only one that does. Under a skewed prior, a mechanism that routes the rare classes onto the labels also emitted by the dominant class can be more ambiguous than uniform labeling: with $C = 3$ and $p = (0.98, 0.01, 0.01)$, the uniform mechanism gives 0.091 bits while the valid mechanism with rows $(0, \frac{1}{2}, \frac{1}{2})$, $(0, 0, 1)$, $(0, 1, 0)$ gives 0.141 bits. "Uniform labeling is maximally ambiguous" is therefore a balanced-prior statement, which is the regime of our benchmarks. The collected BICL matrices violate validity only through their small diagonal noise (Table 7); we never apply Theorem 1 to them, and the results we do apply, Theorem 2 and Corollary 2, do not need validity.

A.2 Proof of Theorem 2

Proof. **Step 1. Bound the entropy of each row.** Fix a class i with $p_i > 0$. The row $Q_{i,:} = (Q_{i1}, \ldots, Q_{iC})$ is the conditional distribution of $\bar{Y}$ given $Y = i$, and by assumption $|\operatorname{supp} Q_{i,:}| \le k$. A distribution supported on at most k points has entropy at most $\log k$ (maximum entropy on a finite support), so

$$H(\bar{Y} \mid Y = i) = H(Q_{i,:}) \le \log k.$$

Step 2. Average over the true class. By the definition of conditional entropy and Step 1,

$$\begin{aligned} H(\bar{Y} \mid Y) &= \sum_{i:\, p_i > 0} p_i \, H(\bar{Y} \mid Y = i) \\ &\le \sum_{i:\, p_i > 0} p_i \log k \\ &= \log k. \end{aligned}$$

Step 3. Convert to a lower bound on the mutual information. Using $I(Y; \bar{Y}) = H(\bar{Y}) - H(\bar{Y} \mid Y)$ together with Step 2,

$$I(Y; \bar{Y}) = H(\bar{Y}) - H(\bar{Y} \mid Y) \;\ge\; H(\bar{Y}) - \log k.$$

Step 4. Convert to an upper bound on the ambiguity. The mutual information also equals $H(Y) - H(Y \mid \bar{Y})$. Solving for $H(Y \mid \bar{Y})$ and applying Step 3,

$$\begin{aligned} H(Y \mid \bar{Y}) &= H(Y) - I(Y; \bar{Y}) \\ &\le H(Y) - \big(H(\bar{Y}) - \log k\big) \\ &= H(Y) + \log k - H(\bar{Y}). \end{aligned}$$

This is identity (1) with $H(\bar{Y} \mid Y)$ replaced by its upper bound $\log k$.

Step 5. Specialize to a balanced prior. If $p_i = 1/C$ then $H(Y) = \log C$. Substituting this and $H(\bar{Y}) = \log C - \delta$, where $\delta = \log C - H(\bar{Y})$, into Step 4 gives

$$H(Y \mid \bar{Y}) \le \log C + \log k - (\log C - \delta) = \log k + \delta.$$

Finally, $\delta \ge 0$ because $H(\bar{Y}) \le \log C$ for any distribution on C points, and, using $\sum_j q_j = 1$ and $H(\bar{Y}) = -\sum_{j:\, q_j > 0} q_j \log q_j$,

$$\delta = \log C - H(\bar{Y}) = \sum_{j:\, q_j > 0} q_j \log C + \sum_{j:\, q_j > 0} q_j \log q_j = \sum_{j:\, q_j > 0} q_j \log \frac{q_j}{1/C} = D_{\mathrm{KL}}\big(q \,\big\|\, \mathrm{Unif}[C]\big).$$

This completes the proof. □

In the bound $\log k + \delta$, the first term measures class-conditional concentration (the most ambiguous k-sparse row is uniform on k labels and has entropy exactly $\log k$), and the second measures the loss of global output diversity relative to a uniform use of the C complementary labels. Theorem 2

thus decomposes ambiguity into two competing effects, a within-class term that sparsity caps and a diversity term that sparsity can worsen; a small k lowers ambiguity only if δ does not grow at the same time. Nothing in the proof depends on the diagonal of Q. The bound is tight: equality holds exactly when every row is uniform on k entries, and two matrices in the paper have this form and sit on it, Q^{sparse} of Appendix A.4 ($\log 4 + 1.053 = 3.053$ bits) and the rule-based annotator of Appendix C.2 ($\log 4 + 0.200 = 2.200$ bits).

A.3 Proofs of Corollaries 1 and 2

Corollary 1 (Sufficient condition for lower ambiguity than uniform labeling)**.** *Under a balanced prior, if every row of Q has support size at most k and $\delta < \log \frac{C-1}{k}$, then $H_Q(Y \mid \bar{Y}) < \log(C-1) = H_{\text{unif}}(Y \mid \bar{Y})$ with margin at least $\log \frac{C-1}{k} - \delta$, i.e., $I_Q(Y;\bar{Y}) \geq \log C - \log k - \delta$.*

Proof of Corollary 1. Theorem 2 under the balanced prior gives the first inequality below, and the hypothesis $\delta < \log \frac{C-1}{k}$ gives the second:

$$H_Q(Y \mid \bar{Y}) \;\leq\; \log k + \delta \;<\; \log k + \log \frac{C-1}{k} \;=\; \log(C-1) \;=\; H_{\text{unif}}(Y \mid \bar{Y}),$$

where the last equality is Step 2 of the proof of Theorem 1. Subtracting the first bound from $\log(C-1)$ gives the margin,

$$\log(C-1) - H_Q(Y \mid \bar{Y}) \;\geq\; \log(C-1) - \log k - \delta \;=\; \log \frac{C-1}{k} - \delta \;>\; 0.$$

For the information bound, use $H(Y) = \log C$ under the balanced prior and the first inequality above:

$$\begin{aligned} I_Q(Y;\bar{Y}) &= \log C - H_Q(Y \mid \bar{Y}) \\ &\geq \log C - \log k - \delta. \end{aligned}$$

□

Corollary 1 assumes exactly k-sparse rows, which the analysis case satisfies by construction with $k = 4$ (on CIFAR-20 two classes carry a fifth label with mass below 1.3%, so $k = 5$ there). Practical BICL does not: after clustering, most rows carry a handful of rare answers, which inflates the maximal row support far beyond four even though almost all of the mass in each row still sits on a few entries. The following variant charges only for the mass that actually lies outside the k_0 largest entries, and it is the form we use for the practical matrices.

Corollary 2 (Robust form of Theorem 2)**.** *For $k_0 < C$ let ε_i be the probability mass of row i of Q outside its k_0 largest entries and $\bar{\varepsilon} = \sum_i p_i \varepsilon_i$. Then $H(\bar{Y} \mid Y) \leq B_{k_0} := h_b(\bar{\varepsilon}) + (1-\bar{\varepsilon}) \log k_0 + \bar{\varepsilon} \log(C-k_0)$, where h_b is the binary entropy function, and under a balanced prior $H(Y \mid \bar{Y}) \leq B_{k_0} + \delta$. When $\bar{\varepsilon} = 0$, $B_{k_0} = \log k_0$ and the bound reduces to that of Theorem 2 with $k = k_0$, from which Corollary 1 follows.*

Proof of Corollary 2. **Step 1. Partition each row into its k_0 largest entries and the rest.** Fix a class i with $p_i > 0$. Let $T_i \subseteq [C]$ be a set of indices of k_0 largest entries of the row $Q_{i,:}$ (ties broken arbitrarily), let $T_i^c = [C] \setminus T_i$, and let

$$\varepsilon_i := \sum_{j \notin T_i} Q_{ij} = 1 - \sum_{j \in T_i} Q_{ij} \in [0,1]$$

be the mass of the row outside T_i. For an index set T on which the row puts positive mass, write $Q_{i,:} \mid T$ for the row restricted to T and renormalized to sum to one.

Step 2. Apply the grouping property of entropy. Drawing $\bar{Y}$ from $Q_{i,:}$ can be carried out in two stages: first decide whether $\bar{Y} \in T_i$ (probability $1-\varepsilon_i$) or $\bar{Y} \in T_i^c$ (probability ε_i), then draw $\bar{Y}$ from the corresponding renormalized part. By the grouping property of entropy, the entropy of the row is the entropy of the first stage plus the expected entropy of the second stage:

$$H(Q_{i,:}) = h_b(\varepsilon_i) + (1-\varepsilon_i)\, H(Q_{i,:} \mid T_i) + \varepsilon_i\, H(Q_{i,:} \mid T_i^c),$$

with the convention that a term whose weight is zero is omitted (if $\varepsilon_i = 0$ the last term is absent).

Step 3. Bound the two conditional entropies by their support sizes. The distribution $Q_{i,:} \mid T_i$ lives on $|T_i| = k_0$ points and $Q_{i,:} \mid T_i^c$ on $|T_i^c| = C - k_0$ points, so by maximum entropy on a finite support

$$H(Q_{i,:} \mid T_i) \le \log k_0, \qquad H(Q_{i,:} \mid T_i^c) \le \log(C - k_0).$$

Substituting these into Step 2 and using that the weights $1 - \varepsilon_i$ and ε_i are nonnegative,

$$H(Q_{i,:}) \le h_b(\varepsilon_i) + (1 - \varepsilon_i) \log k_0 + \varepsilon_i \log(C - k_0).$$

Step 4. Average over the true class. Multiplying by p_i and summing over i,

$$\begin{aligned} H(\bar{Y} \mid Y) &= \sum_i p_i \, H(Q_{i,:}) \\ &\le \sum_i p_i \, h_b(\varepsilon_i) + \log k_0 \sum_i p_i(1 - \varepsilon_i) + \log(C - k_0) \sum_i p_i \varepsilon_i \\ &= \sum_i p_i \, h_b(\varepsilon_i) + (1 - \bar{\varepsilon}) \log k_0 + \bar{\varepsilon} \log(C - k_0), \end{aligned}$$

where $\bar{\varepsilon} = \sum_i p_i \varepsilon_i$ and we used $\sum_i p_i = 1$. The last two terms pass through the average exactly because they are linear in ε_i.

Step 5. Apply the Jensen inequality to the binary-entropy term. The binary entropy h_b is concave on $[0, 1]$ and $(p_i)_i$ is a probability vector, so the Jensen inequality gives

$$\sum_i p_i \, h_b(\varepsilon_i) \le h_b\Big(\sum_i p_i \varepsilon_i \Big) = h_b(\bar{\varepsilon}).$$

Combining this with Step 4,

$$H(\bar{Y} \mid Y) \le h_b(\bar{\varepsilon}) + (1 - \bar{\varepsilon}) \log k_0 + \bar{\varepsilon} \log(C - k_0) = B_{k_0}.$$

Step 6. Convert to the ambiguity bound. Under the balanced prior, $H(Y) = \log C$ and $H(\bar{Y}) = \log C - \delta$, so identity (1) reads

$$H(Y \mid \bar{Y}) = H(Y) + H(\bar{Y} \mid Y) - H(\bar{Y}) = H(\bar{Y} \mid Y) + \delta,$$

and therefore, by Step 5,

$$H(Y \mid \bar{Y}) \le B_{k_0} + \delta.$$

Finally, if $\bar{\varepsilon} = 0$ then every row is supported on its k_0 largest entries, $h_b(0) = 0$, and $B_{k_0} = \log k_0$, which is Theorem 2 with $k = k_0$ and hence yields Corollary 1. This completes the proof. □

Remarks. Three points clarify what the corollaries add, and a fourth delimits their scope. First, for a balanced prior and a valid Q, the uniqueness part of Theorem 1 already implies that every non-uniform mechanism, and hence every k-sparse one with $k < C - 1$, is strictly less ambiguous than uniform labeling; what Corollary 1 adds is a quantitative margin, $\log \frac{C-1}{k} - \delta$, that survives without validity and therefore applies to matrices with diagonal noise. Second, the condition $\delta < \log \frac{C-1}{k}$ is sufficient but not necessary: in a check of 3,000 random k-sparse mechanisms with $C \in \{5, 10, 20, 50, 100\}$, k drawn from $\{1, \ldots, \min(C - 1, 12)\}$ and Gamma-distributed row weights, the ambiguity was below $\log(C - 1)$ every time, while the condition failed in 290 of them (205 with $k = C - 1$, where it cannot hold, and 85 with $k < C - 1$). Third, there is an exact criterion. With the effective row support $k_{\text{eff}} := 2^{H(\bar{Y}|Y)} \le k$, identity (1) under a balanced prior reads $H(Y \mid \bar{Y}) = \log k_{\text{eff}} + \delta$, so a mechanism is less ambiguous than uniform labeling if and only if $\delta < \log \frac{C-1}{k_{\text{eff}}}$; Table 7 reports k_{eff} for every mechanism. Finally, on scope: a lower $H(Y \mid \bar{Y})$ does not by itself order the test errors of CLL algorithms, which also depend on the model class, the optimizer, the estimate of Q and the sample size, and it is not identifiability, which asks for distinct rows of Q rather than low ambiguity. Theorem 1 concerns the idealized valid mechanism, while the results we apply to the collected matrices do not need $Q_{ii} = 0$.

A.4 Why Sparsity Alone Is Not Enough

It would be convenient if a sparser Q were always a more informative one. It is not, and a small 10-class example shows why. Start from the dense biased matrix

$$Q^{\text{dense}} = \begin{pmatrix} 0 & .02 & .02 & .02 & .02 & .90 & .005 & .005 & .005 & .005 \\ .02 & 0 & .02 & .02 & .02 & .005 & .90 & .005 & .005 & .005 \\ .02 & .02 & 0 & .02 & .02 & .005 & .005 & .90 & .005 & .005 \\ .02 & .02 & .02 & 0 & .02 & .005 & .005 & .005 & .90 & .005 \\ .02 & .02 & .02 & .02 & 0 & .005 & .005 & .005 & .005 & .90 \\ .02 & .02 & .02 & .02 & .90 & 0 & .005 & .005 & .005 & .005 \\ .02 & .02 & .02 & .02 & .005 & .90 & 0 & .005 & .005 & .005 \\ .02 & .02 & .02 & .02 & .005 & .005 & .90 & 0 & .005 & .005 \\ .02 & .02 & .02 & .02 & .005 & .005 & .005 & .90 & 0 & .005 \\ .02 & .02 & .02 & .02 & .005 & .005 & .005 & .005 & .90 & 0 \end{pmatrix},$$

and form a 4-sparse matrix by keeping four entries in every row, zeroing the rest, and re-normalizing:

$$Q^{\text{sparse}} = \begin{pmatrix} 0 & .25 & .25 & .25 & .25 & 0 & 0 & 0 & 0 & 0 \\ .25 & 0 & .25 & .25 & .25 & 0 & 0 & 0 & 0 & 0 \\ .25 & .25 & 0 & .25 & .25 & 0 & 0 & 0 & 0 & 0 \\ .25 & .25 & .25 & 0 & .25 & 0 & 0 & 0 & 0 & 0 \\ .25 & .25 & .25 & .25 & 0 & 0 & 0 & 0 & 0 & 0 \\ .25 & .25 & .25 & .25 & 0 & 0 & 0 & 0 & 0 & 0 \\ .25 & .25 & .25 & .25 & 0 & 0 & 0 & 0 & 0 & 0 \\ .25 & .25 & .25 & .25 & 0 & 0 & 0 & 0 & 0 & 0 \\ .25 & .25 & .25 & .25 & 0 & 0 & 0 & 0 & 0 & 0 \\ .25 & .25 & .25 & .25 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Under a balanced prior, decomposition (1) gives the values in Table 6.

Table 6: What sparsification does to the example: both terms of (1) move against information, and the ambiguity $H(Y \mid \bar{Y})$ rises by 1.86 bits. Values in bits under a balanced prior; the last column counts complementary labels that never occur.

	$H(\bar{Y} \mid Y)$	$H(\bar{Y})$	δ	$H(Y \mid \bar{Y})$	$I(Y;\bar{Y})$	empty cols
Q^{dense}	0.741	2.866	0.456	1.197	2.124	0
Q^{sparse}	2.000	2.269	1.053	3.053	0.269	5
change	$+1.259$	-0.597	$+0.597$	$+1.855$	-1.855	

Sparsification makes the mechanism markedly *more* ambiguous, by 1.86 bits, and the reason is visible in both terms of (1). The retained rows become flatter: the surviving entries are the small ones, so after re-normalization each row is uniform on four labels and attains the largest possible 4-sparse within-class entropy, whereas each dense row was dominated by a single entry. At the same time the outputs collapse: every row is mapped onto the same four columns, so half of the complementary labels never occur and the global diversity term falls. The two effects add, and their sum is the observed increase. Which entries are retained is decisive. Keeping the 0.90 entry and three of the four 0.02 entries of each row instead (its four largest entries, the tie among the 0.02 entries broken arbitrarily) lowers the ambiguity to between 0.93 and 0.98 bits, and keeping only the largest entry to 0.80 bits, both well below the dense value. What matters is therefore not the number of nonzero entries alone, but which transitions are retained and whether they preserve diversity across the output label space. This is precisely the content of Theorem 2: the cap $\log k$ helps only when the dense mechanism did not already beat it, and only when it comes with a small deficit δ. For completeness, Q^{sparse} still satisfies Corollary 1 and is less ambiguous than uniform labeling (3.05 against 3.17 bits); the example concerns the comparison with a structured dense mechanism, which none of our results orders.

A Monte Carlo comparison does not settle the question either. Drawing dense matrices with i.i.d. $U[0,1]$ entries and sparsifying them to four entries per row makes the sparse matrices win in essentially every trial, but such random dense rows are nearly uniform (in our replication

$H(\bar{Y} \mid Y) \approx \log(C-1) - 0.28$ bits for $C \in \{10, 100, 200\}$), so any 4-sparse row, with at most 2 bits, beats them trivially. The simulation illustrates Theorem 2 but says nothing about semantically structured matrices, and we do not use it as evidence.

A.5 Posterior Support and the Analysis Case

Theorem 2 looks at the rows of Q. Looking at its columns instead explains a regularity of the analysis case that is otherwise puzzling: its ambiguity barely changes between 10 and 200 classes.

Proposition 1 (Posterior-support bound). *For each complementary label j let $d_j = |\{i : p(Y = i, \bar{Y} = j) > 0\}|$ (so $d_j \geq 1$ if and only if $q_j > 0$). Then $H(Y \mid \bar{Y} = j) \leq \log d_j$ for every j with $q_j > 0$, and*

$$H(Y \mid \bar{Y}) \;\leq\; \sum_{j:\, q_j>0} q_j \log d_j \;\leq\; \log\Big(\sum_{j:\, q_j>0} q_j d_j\Big) \;\leq\; \log \max_j d_j.$$

Proof. Labels with $q_j = 0$ have $d_j = 0$; they receive zero weight in $H(Y \mid \bar{Y})$ and are excluded from the sums, which is why the sums run over $\{j : q_j > 0\}$.

Step 1. Bound the posterior entropy by the posterior support. Fix j with $q_j > 0$. The posterior $p(Y = i \mid \bar{Y} = j) = p(Y = i, \bar{Y} = j)/q_j$ is positive exactly for the d_j classes with $p(Y = i, \bar{Y} = j) > 0$ (and $d_j \geq 1$ because $q_j > 0$), so it is a distribution on d_j points, and by maximum entropy on a finite support

$$H(Y \mid \bar{Y} = j) \leq \log d_j.$$

Step 2. Average over $\bar{Y}$. By the definition of conditional entropy and Step 1,

$$\begin{aligned} H(Y \mid \bar{Y}) &= \sum_{j:\, q_j>0} q_j \, H(Y \mid \bar{Y} = j) \\ &\leq \sum_{j:\, q_j>0} q_j \log d_j. \end{aligned}$$

Step 3. Apply the Jensen inequality to the logarithm. The logarithm is concave on $(0, \infty)$ and $(q_j)_{j:\, q_j>0}$ is a probability vector, so the Jensen inequality applied to the positive values d_j gives

$$\sum_{j:\, q_j>0} q_j \log d_j \leq \log\Big(\sum_{j:\, q_j>0} q_j d_j\Big).$$

Since a weighted average of the d_j cannot exceed their maximum and $\log$ is increasing,

$$\log\Big(\sum_{j:\, q_j>0} q_j d_j\Big) \leq \log \max_j d_j.$$

Chaining Steps 2 and 3 gives the three inequalities of the proposition. □

The mechanism is simple. In the analysis case each class i receives a candidate set S_i of k labels drawn uniformly from $[C] \setminus \{i\}$ and the annotator answers from S_i, so $\operatorname{supp} Q_{i,:} \subseteq S_i$. A given complementary label j can therefore be produced only by the classes whose candidate set contains it, so $d_j \leq \tilde{d}_j := |\{i \neq j : j \in S_i\}|$, and because each class $i \neq j$ includes j in its set independently with probability $\frac{k}{C-1}$, $\tilde{d}_j$ is $\mathrm{Binomial}(C-1, \frac{k}{C-1})$ with mean k, independently of C. The posterior support, and with it the bound of Proposition 1 (which is increasing in the d_j), is thus governed by k rather than by the size of the label space. (If the annotator were uniform within S_i, every allowed class would produce j, so $d_j = \tilde{d}_j$ and $q_j = d_j/(Ck)$, giving $\mathbb{E}\sum_j q_j d_j = k + 1 - \frac{k}{C-1}$; by the Jensen inequality the expected value of the bound $\log \sum_j q_j d_j$ is then at most $\log(k + 1 - \frac{k}{C-1})$, between 2.19 and 2.32 bits for $k = 4$ and C from 10 to 200. The rule-based annotator of Appendix C.2 has $H(Y \mid \bar{Y}) = 2.20$ bits on CIFAR-20, equal to $\sum_j q_j \log d_j$ because its posteriors are uniform, with $\log \sum_j q_j d_j = 2.29$.) The collected analysis-case matrices behave exactly this

way: their mean column support stays close to $k = 4$ on all four datasets, and $H(Y \mid \bar{Y})$ stays within 1.78–1.89 bits from $C = 10$ to $C = 200$ (Table 7).

Clustering weakens this property. A cluster contains several true classes, all of which inherit its candidate set, so a complementary label can now be produced by many classes: the mean column support grows from about four to between 9 and 51 across the four datasets, and the posterior-support bound becomes vacuous for $C \leq 20$ and loose for $C \geq 100$. The lost diversity also shows up in δ. On TinyImageNet-200, 93 of the 200 complementary labels never occur in the practical matrix, which alone costs $\log \frac{200}{107} = 0.90$ bits; with the uneven use of the remaining labels, $\delta = 1.39$ bits, which is why that matrix carries less information than the CIFAR-100 one despite the larger label space. This is the price of clustering relative to the oracle design, and it explains why practical-case ambiguity grows with C while analysis-case ambiguity does not.

A.6 Empirical Information Statistics

Computation. For every mechanism we build the joint distribution $J_{ij} = p_i Q_{ij}$ from a class prior p and a transition matrix Q, and compute $H(Y) = H(p)$, $H(\bar{Y}) = H(q)$ with $q_j = \sum_i J_{ij}$, $H(\bar{Y} \mid Y) = \sum_i p_i H(Q_{i,:})$, $H(Y \mid \bar{Y}) = \sum_j q_j H(J_{:,j}/q_j)$, and $I(Y;\bar{Y})$ from the joint. The computation verifies the identities $I(Y;\bar{Y}) = H(Y) - H(Y \mid \bar{Y}) = H(\bar{Y}) - H(\bar{Y} \mid Y) = H(Y) + H(\bar{Y}) - H(Y,\bar{Y})$ to within 10^{-9} bits for every reported row. We never infer posterior uniformity from row-wise properties of Q; every conditional entropy is computed from the induced joint distribution with the stated prior.

Class prior. For all synthetic mechanisms (uniform, biased) and for all BICL matrices we use the balanced prior $p_i = 1/C$. This is the prior of the benchmarks and of the training labels of Section 5, where complementary labels are sampled from the stored Q for every training image. For the released CLImage and ACLImage datasets we use the empirical class prior of the released training set and the first complementary label of each image. The CIFAR-20 release is mildly imbalanced, with 2,000–3,000 images per superclass; under a balanced prior its values change by at most 0.004 bits.

Estimation considerations. Q for a BICL dataset is the row-normalized count matrix of (true class, complementary label) pairs over the annotated training images: at most 5,000, 2,500, and 500 images per class on CIFAR-10/20/100, with a few classes slightly below that, while on TinyImageNet-200 the annotated pool per class is not uniform and cannot be recovered from the row-normalized matrix. All reported values are plug-in quantities of these matrices, and plug-in entropies from finite counts are biased downward by roughly $(m-1)/(2N \ln 2)$ bits for a row of support m estimated from N samples. This bias matters in one place. For the uniform reference, a sampled matrix with 500 images per class would understate $H(Y \mid \bar{Y})$ by 0.15–0.35 bits at $C = 100$ and $C = 200$ and overstate the tiny value of $I(Y;\bar{Y})$ by an order of magnitude, so we report the analytic values $\log(C-1)$ and $\log \frac{C}{C-1}$ instead. For the BICL matrices the corresponding Miller–Madow correction is at most $+0.01$ bits on CIFAR-10/20 and about $+0.07$ bits on CIFAR-100, too small to alter any comparison in Table 7.

Label noise. The collected matrices $\hat{Q}^{\mathrm{BICL}}$ carry a small diagonal mass equal to the label-noise rate (column "noise" in Table 7). We leave it in place: all statistics are computed on the matrices as measured, and the only results we apply to them, Theorem 2 and Corollary 2, do not require $Q_{ii} = 0$. Theorem 1 is a statement about the idealized valid mechanism.

Empirical interpretation. Three patterns stand out in Table 7. First, uniform and uniform-like annotation remains highly ambiguous: the human-annotated CLImage matrices are in fact slightly more ambiguous than the analytic uniform reference, and the VLM-annotated uniform designs, including the same design at 100 and 200 classes, stay within a few tenths of a bit of it, so neither annotation noise nor VLM preferences move a uniform design far from the uniform mechanism. Second, the analysis-case BICL mechanism preserves information as C grows, with ambiguity flat at about 1.8 bits from 10 to 200 classes, in line with Appendix A.5. Third, practical BICL loses part of this advantage to clustering, and the loss grows with C, yet it remains substantially more informative than every uniform-like mechanism on every benchmark. Taken together, these observations suggest

that the information advantage of BICL comes mainly from the structure induced by candidate-set construction rather than from the VLM, and that clustering quality is what separates the deployable protocol from the oracle design. These are properties of the supervision mechanisms; whether a learning algorithm exploits them is the empirical question of Section 5.

Table 7: Information statistics of every complementary-label mechanism in the paper, in bits. The prior is balanced except for the two released datasets marked †, where we use their empirical class prior and the first complementary label of each image. k is the largest row support, $k_{\text{eff}} = 2^{H(\bar{Y}|Y)}$, $\bar{d}$ the mean column support, and noise the diagonal mass in percent. The last two columns check the condition of Corollary 1, $\delta < \log \frac{C-1}{k}$, and report the bound $B_4 + \delta$ of Corollary 2, with a check mark when it falls below $\log(C-1)$. BICL-A and BICL are the analysis and practical cases; Uniform (VLM) is the ACLImage-style uniform design annotated with Qwen3-8B. All values are computed directly from the stored matrices (Appendix A.6).

Dataset	Mechanism	k	k_{eff}	$\bar{d}$	noise	$H(\bar{Y} \mid Y)$	$H(Y \mid \bar{Y})$	$I(Y;\bar{Y})$	δ	Cor. 1	$B_4 + \delta$
CIFAR-10	Uniform (analytic)	9	9.0	9.0	0.00	3.170	3.170	0.152	0.000	–	–
	Biased (Tab. 1)	9	5.5	9.0	0.00	2.457	2.509	0.813	0.052	–	2.809✓
	CLImage (human)†	10	9.3	10.0	3.98	3.223	3.240	0.082	0.018	–	3.278
	ACLImage (VLM)†	10	7.7	9.7	0.21	2.953	3.019	0.303	0.067	–	3.189
	BICL-A	4	2.9	4.0	0.00	1.526	1.803	1.518	0.278	✓	2.278✓
	BICL	10	3.7	8.7	0.23	1.880	2.200	1.122	0.320	–	2.889✓
CIFAR-20	Uniform (analytic)	19	19.0	19.0	0.00	4.248	4.248	0.074	-0.000	–	–
	Biased (Tab. 1)	19	5.6	19.0	0.00	2.478	2.599	1.722	0.122	–	3.119✓
	CLImage (human)†	20	18.7	20.0	2.32	4.224	4.267	0.052	0.045	–	4.327
	ACLImage (VLM)†	20	14.9	19.4	0.88	3.894	4.198	0.121	0.306	–	4.430
	BICL-A	5	2.4	3.8	0.00	1.291	1.782	2.540	0.490	✓	2.492✓
	BICL	19	7.4	16.5	0.80	2.884	3.618	0.704	0.734	–	4.090✓
CIFAR-100	Uniform (analytic)	99	99.0	99.0	0.00	6.629	6.629	0.014	-0.000	–	–
	Uniform (VLM)	97	75.5	91.7	0.07	6.239	6.424	0.220	0.185	–	6.750
	BICL-A	4	2.7	3.9	0.00	1.445	1.826	4.818	0.381	✓	2.381✓
	BICL	68	17.2	48.9	0.03	4.100	4.573	2.071	0.472	✓	5.585✓
TinyImageNet-200	Uniform (analytic)	199	199.0	199.0	0.00	7.637	7.637	0.007	0.000	–	–
	Uniform (VLM)	182	142.0	169.0	0.05	7.150	7.277	0.367	0.127	–	7.720
	BICL-A	4	2.7	3.9	0.00	1.432	1.889	5.755	0.457	✓	2.457✓
	BICL	78	21.9	50.6	0.05	4.453	5.839	1.805	1.386	–	7.376✓

B Related Work

CL and MCL. The seminal work of Ishida et al. (2017) introduced CLL under a uniform complementary-label assumption. Later, Ishida et al. (2019) proposed a general unbiased risk estimator along with non-negative and gradient-ascent variants to mitigate negative empirical risk and overfitting. Chou et al. (2020) proposed the surrogate complementary loss (SCL) framework that aligns empirical gradients with the true gradients to mitigate overfitting. Besides complementary label (CL), multiple complementary-label learning (MCL) allows collecting several CLs per instance. Feng et al. (2020) formalized MCL and proposed an unbiased risk estimator to minimize all MCLs of each instance altogether. A recent work connects CLL to other weakly supervised paradigms to improve representations (Jiang et al., 2024).

Biased and noisy CL. Yu et al. (2018) modeled biased complementary-label generation with a transition matrix and introduced forward correction (FWD). Subsequent studies targeted the estimation of the transition matrix. The Selected-Completely-at-Random (SCARCE) Wang et al. (2024b) removed the need for a uniform complementary-label distribution and avoided requiring an auxiliary ordinary-labeled set for transition-matrix estimation. Lin & Lin (2023) quantified the effect of an inaccurate transition matrix. The complementary one-versus-rest (COVR) loss further addresses biased distributions by increasing the complementary logit margin to improve robustness You et al. (2025). Ishiguro et al. (2022) developed a noise-robust CLL framework by modeling the noise via a noise transition matrix and analyzing consistency under matrix misspecification.

Benchmarks, toolkits, and class-imbalance. Wang et al. (2025) and Mai et al. (2025) released real-world datasets (CLImage and ACLImage), highlighting practical challenges such as noise, bias, and imbalanced annotations. To improve reproducibility and fair comparison, Ye et al. (2024) released

`libcll`, a Python toolkit that standardizes datasets, assumptions, and algorithm implementations under a unified interface. Wei et al. (2023) addressed class-imbalanced training data via weighted empirical risk minimization and provided supporting generalization analysis.

C Further Analysis and Discussion

This appendix provides further analyses to better understand the behavior and properties of BICL beyond the main experimental results. We examine the supervision quality induced by BICL, identify the source of its performance improvement, and study the effects of class imbalance and key design choices. We further analyze the number of clusters and candidate labels, statistical significance, and the characteristics of the resulting BICL datasets; limitations and future directions are discussed in Section 6.

C.1 Supervision Quality of the BICL Datasets

Table 7 in Appendix A.6 reports the conditional entropy $H(Y \mid \bar{Y})$ and the mutual information $I(Y;\bar{Y})$ of every complementary-label mechanism in the main comparison, computed from the stored transition matrices with a single verified routine. Among the compared datasets, the BICL matrices have the lowest ambiguity and carry the most information about the true class on every benchmark. These are properties of the supervision mechanism; whether a given learning algorithm turns the additional information into accuracy is an empirical question, which Section 5 and the ablations below address. Figure 4 shows the test accuracy during training for the main comparison: BICL reaches its peak performance faster than Uniform, CLImage, and ACLImage, reaching strong performance in fewer training epochs.

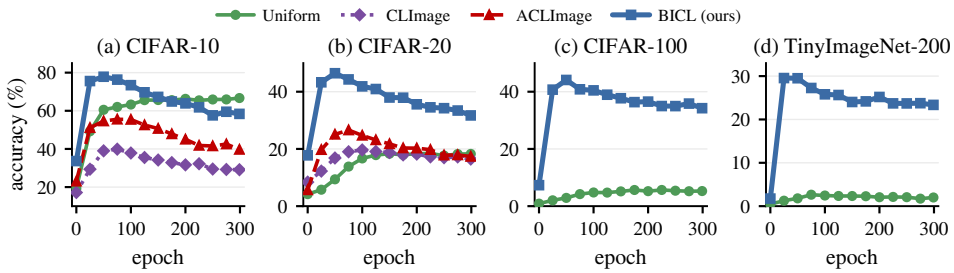


Figure 4: Test accuracy during training on four datasets. Our method reaches its peak performance faster than Uniform, CLImage, and ACLImage when applicable. CLImage and ACLImage are limited to at most 20 classes, so panels (c) and (d) omit them.

C.2 Ablation on the Source of Improvement in BICL: Label Selection Bias vs. VLM Prior Knowledge

To examine where the improvement of BICL comes from, the structured transition or the prior knowledge of the VLM annotator, we run an ablation with two annotators:

1. **VLM annotator**: the setting of Appendix C.5 with four candidate labels per class, where the true label is excluded from the candidate set.
2. **Rule-based annotator**: four candidate labels per class, none equal to the true class, from which one is selected uniformly at random; this is possible because the analysis case (Figure 2) has access to the true class.

The results of (1) and (2) in Table 8 are similar, which suggests that the preferences of the VLM itself are not the main driver of the gain. Instead, the improvement appears to come from the structured transition mechanism that the constrained candidate sets induce, which both annotators share. We note that the collected CLs are instance-dependent, so the class-conditional matrix Q is only a marginal model of the annotation process; that two annotators with four-candidate sets, and hence

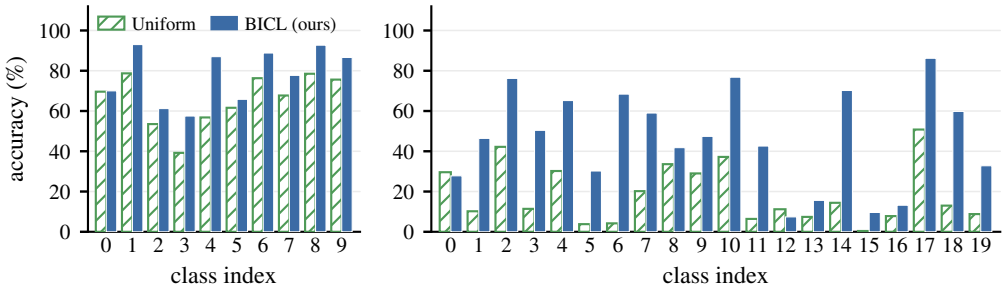
Table 8: Comparison between the VLM annotator and the rule-based annotator on CIFAR-20 (accuracy (%), mean±std).

Annotator	Method	Dataset	Accuracy (%)
(1) VLM annotator	FWD	CIFAR-20	68.89±0.87
	CPE-F	CIFAR-20	71.13±1.00
(2) A rule-based annotator	FWD	CIFAR-20	69.80±0.79
	CPE-F	CIFAR-20	70.29±1.24

transition matrices of similar structure, reach similar accuracy is evidence that this marginal structure captures much of what the learner exploits, not a guarantee that it captures everything. Table 5 in Section 5 repeats the comparison in the practical case, on the collected clusters and candidate sets of CIFAR-20 and CIFAR-100, with the same conclusion.

C.3 Effect of Biased Data Collection on Minority-Class Performance

To better understand the effect of the imbalance induced by BICL, we conduct an ablation study to compare the per-class accuracy under a uniform complementary-label distribution, where each complementary class is balanced, with that under the BICL distribution, as shown in the following Figure 5.



(a) CIFAR-10 (b) CIFAR-20

Figure 5: Per-class accuracy under the uniform complementary-label distribution (green hatched outline) and the BICL distribution (blue) on CIFAR-10 (left) and CIFAR-20 (right) with the FWD algorithm. Although BICL induces an imbalanced complementary-label distribution, it does not degrade performance on minority classes; it is better than the uniform distribution on every CIFAR-10 class and on 18 of the 20 CIFAR-20 classes, and on classes on which uniform CLL collapses to a few percent.

These results suggest that the imbalanced complementary-label distribution induced by BICL does not hurt per-class accuracy; on these datasets it is in fact higher on almost every class. This highlights that the imbalance in CLL, especially in the complementary-label space, behaves differently from class imbalance in ordinary label learning, where minority classes are often severely affected.

C.4 The Optimal Number of Clusters

In the label selection phase, the granularity of K-means clustering, determined by the number of clusters Z_c, is a key hyperparameter for grouping visually similar images prior to VLM annotation. We conduct an ablation study to examine the sensitivity of BICL to the number of clusters used in the label-selection stage (Figure 1). Specifically, we vary the cluster count while keeping all other components of BICL unchanged. For each setting, the entire BICL pipeline is rerun, including feature clustering and re-annotation of CLs by the VLM under the newly induced candidate label sets. Although this procedure is computationally expensive, it allows for a controlled evaluation of how the cluster granularity affects performance.

We explore cluster sizes ranging from $|C|$ to $4|C|$, where $|C|$ denotes the number of classes in the dataset. The results in Table 9 show that setting the number of clusters equal to the class cardinality ($|C|$) yields the best or near-best performance across all four datasets. Increasing the number of clusters beyond $|C|$ consistently degrades performance, particularly on large-scale datasets such as CIFAR-100 and TinyImageNet-200.

This trend can be attributed to the trade-off between semantic coherence and granularity in the clustering process. When the number of clusters matches the number of classes, clusters tend to capture class-level structure, leading to more informative and stable candidate label sets for complementary labeling. In contrast, using a larger number of clusters fragments semantically similar samples into smaller groups, increasing annotation noise and reducing the effectiveness of the induced bias. From a practical perspective, choosing $|C|$ clusters also offers a favorable balance between accuracy and computational cost, making it a natural and robust default choice for BICL.

Table 9: Impact of cluster size on model performance across four datasets. $|C|$ denotes the original number of classes in each dataset.

Dataset	$\|C\|$	$1.5\|C\|$	$2\|C\|$	$3\|C\|$	$4\|C\|$
CIFAR-10	**81.23**$\pm$0.05	80.80$\pm$0.39	78.68$\pm$0.46	79.01$\pm$0.18	77.34$\pm$0.26
CIFAR-20	50.84$\pm$0.32	**51.22**$\pm$0.25	50.52$\pm$0.11	47.95$\pm$0.28	47.52$\pm$0.37
CIFAR-100	**46.70**$\pm$0.60	36.98$\pm$0.40	36.72$\pm$0.33	36.25$\pm$0.17	34.15$\pm$0.64
TinyImageNet-200	**32.15**$\pm$0.30	31.79$\pm$0.23	32.10$\pm$0.23	30.78$\pm$0.40	30.45$\pm$0.31

C.5 Effect of the Number of Sampled Labels

We conduct an ablation study to investigate how the number of sampled candidate labels affects the performance of BICL and to justify our choice of using a fixed number of four labels per instance for VLM annotation. Specifically, we vary the number of sampled labels n from 2 to 10 for each class. Similar to the ablation in Section C.4, the entire BICL pipeline is rerun for each setting: we resample the candidate label sets, request the VLM to re-annotate CLs under the new label space, and retrain the downstream classifier. Although this procedure is computationally expensive, it enables a controlled evaluation of the impact of label-space size on complementary-label quality and downstream performance.

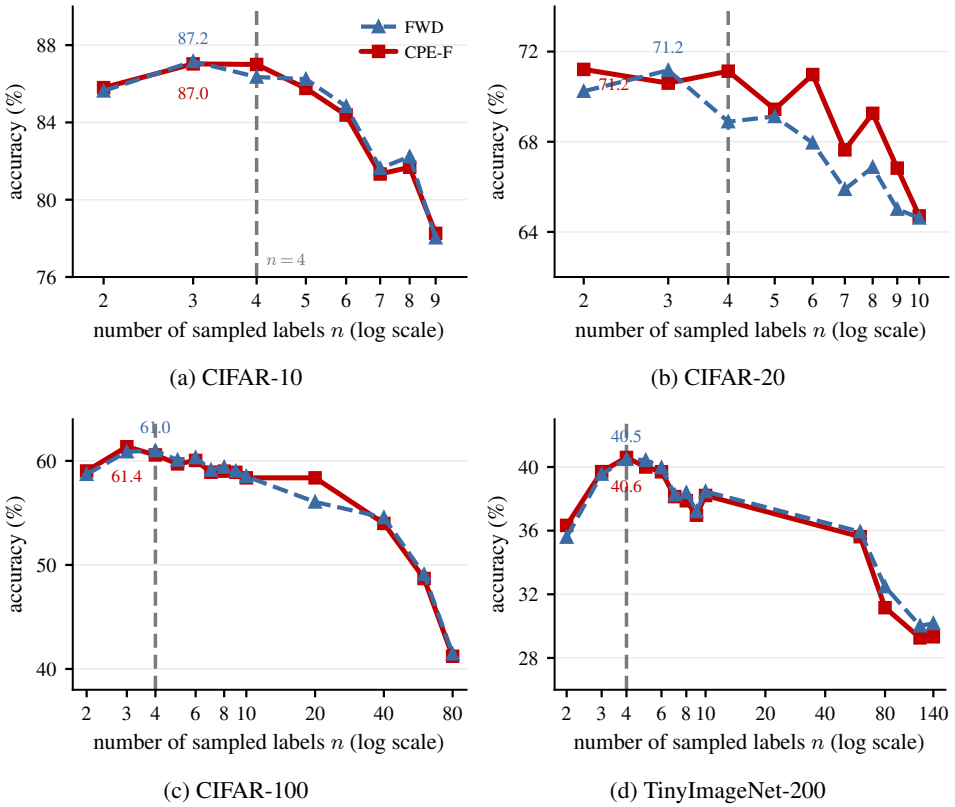
The results, illustrated in Figure 6, show that sampling a small number of labels ($n = 3$ or $n = 4$) consistently yields the best or near-best performance across all datasets and algorithms. In contrast, increasing n beyond this range leads to a steady degradation in accuracy. This trend becomes particularly pronounced as n approaches the Uniform setting, where the full label space is used. These findings help explain why the Uniform assumption performs poorly in practice: presenting too many candidate labels increases ambiguity for the VLM, resulting in noisier and less informative complementary supervision. Between $n = 3$ and $n = 4$, we adopt $n = 4$ in our main experiments. This choice follows prior work (Mai et al., 2025; Wang et al., 2025; Wei et al., 2022), which also uses four candidate labels per instance, ensuring a fair comparison. Moreover, $n = 4$ provides a robust balance between reducing label uncertainty and maintaining sufficient diversity in the candidate set, making it a practical and effective default for BICL.

C.6 Statistical Significance of Experimental Results

To verify the performance improvements of our proposed method BICL against the uniform baseline assumption, we conducted paired t-tests. We performed this analysis across four representative algorithms: FWD, CPE-F, CPE-T, and CPE-I for each dataset.

The experiments were repeated three times with different random seeds. For the significance test, we adopted a confidence threshold $\alpha = 0.05$. We set up all the hypotheses as follows:

- **Null Hypothesis (H_0):** The mean accuracy difference between BICL and the baseline (uniform distribution) is less than or equal to zero.
- **Alternative Hypothesis:** BICL achieves a higher mean accuracy than the baseline.



(a) CIFAR-10 (b) CIFAR-20

(c) CIFAR-100 (d) TinyImageNet-200

Figure 6: BICL accuracy against the number of sampled candidate labels n presented to the VLM, for FWD (blue, dashed, triangles) and CPE-F (red, solid, squares). The grey dashed line marks $n = 4$, the setting used in the main experiments, and the peak of each series is labeled. Accuracy is best or near-best at $n = 3$ or 4 and declines steadily as n grows toward the full label set.

Table 10: Statistical significance (p-values) of the performance improvements of BICL compared to the baseline across datasets. The p-values are calculated using a one-tailed paired t-test with 3 different random seeds. In all cases, $p < 0.05$, confirming the significance of the results.

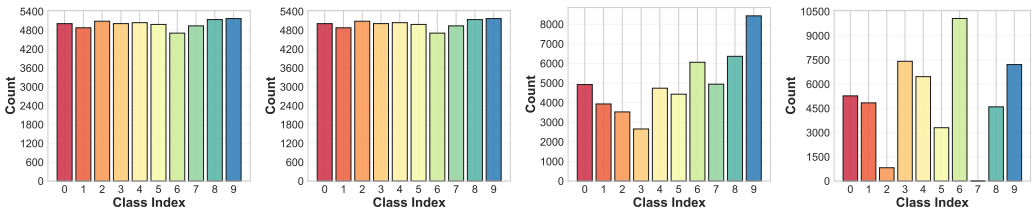
Algorithm	**CIFAR-10**	**CIFAR-20**	**CIFAR-100**	**TinyImageNet-200**
FWD	8.13×10^{-4}	1.41×10^{-4}	3.42×10^{-4}	3.59×10^{-4}
CPE-F	1.9×10^{-3}	9.84×10^{-5}	3.36×10^{-4}	3.75×10^{-4}
CPE-I	3.13×10^{-3}	3.84×10^{-5}	4.32×10^{-5}	3.59×10^{-5}
CPE-T	3.88×10^{-2}	3.82×10^{-4}	6.18×10^{-4}	4.74×10^{-4}

As shown in Table 10, the calculated p-values from all evaluated settings are substantially lower than 0.05. Based on this result, we firmly reject the null hypothesis in favor of the alternative hypothesis. This statistical analysis confirms that the performance gains achieved by BICL reported in Section 5 of the main text are significant.

C.7 Extended BICL Dataset Analysis

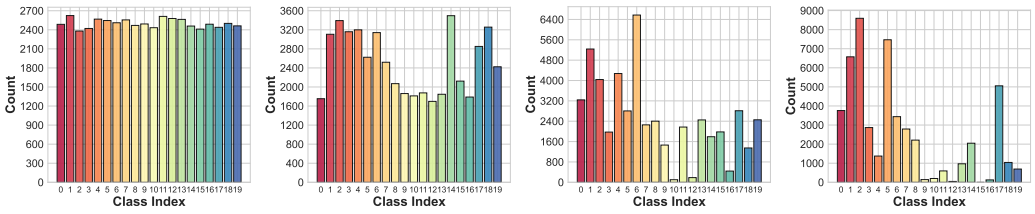
This section provides additional information about the remaining BICL datasets. We validate that the three key characteristics observed in CIFAR-10, biased transitions, label imbalance, and low noise rates, are consistent across datasets with varying numbers of classes and complexity.

As illustrated in Figure 10, Figure 17 the transition matrix for BCLCIFAR-20, BCLCIFAR-100 exhibit distinct, non-uniform patterns similar to the BCLCIFAR-10 results in the main text. This



(a) CIFAR-10 (b) CLCIFAR-10 (c) ACLCIFAR-10 (d) BCLCIFAR-10

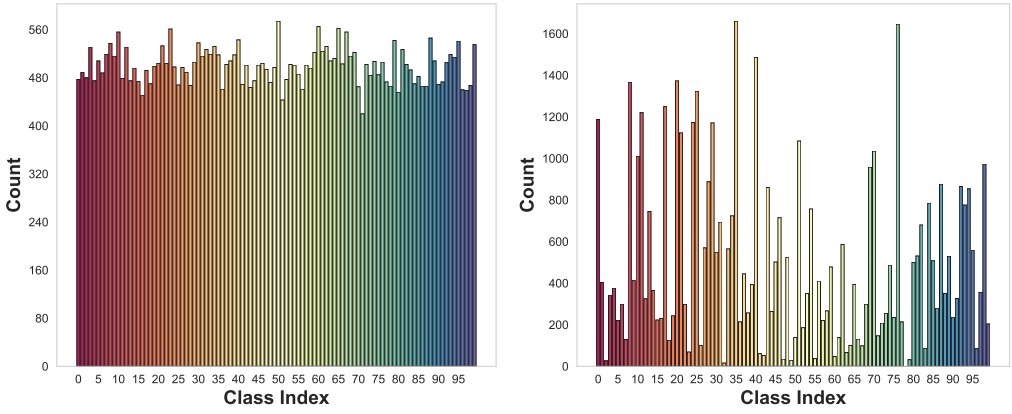
Figure 7: **Label distributions across CIFAR-10 variants**. CIFAR-10 represents an idealized setting with noiseless, uniformly distributed labels. CLCIFAR-10 corresponds to a human-annotated dataset, and ACLCIFAR-10 is VLM-annotated under a uniform distribution design.



(a) CIFAR-20 (b) CLCIFAR-20 (c) ACLCIFAR-20 (d) BCLCIFAR-20

Figure 8: Label distributions across CIFAR-20 variants.

resulted from our design, where we deliberately constrain the labeling space, combined with innate inductive biases characteristic of VLMs.



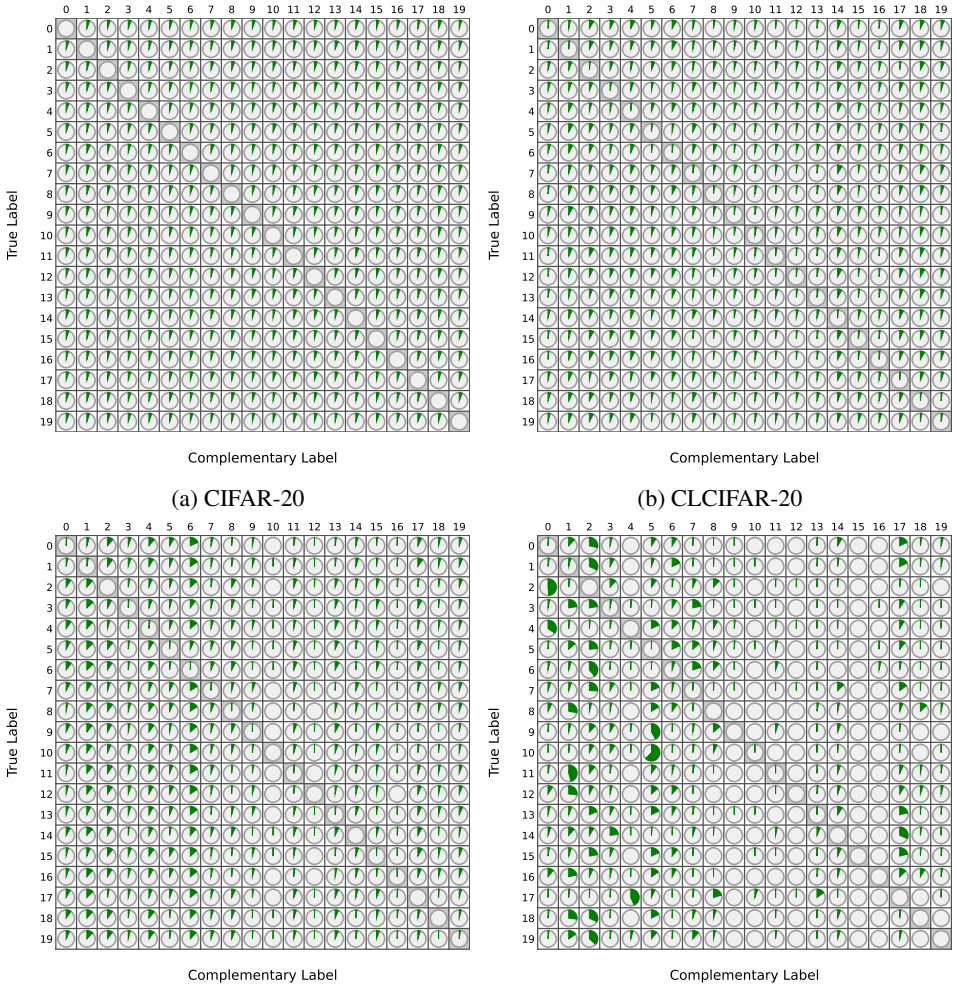
(a) CIFAR-100 (b) BCLCIFAR-100

Figure 9: Label distributions across CIFAR-100 variants.

The imbalanced label distribution is present throughout our datasets.

Low label noise rate. Despite its pronounced bias and imbalance, our dataset maintains a low label noise rate. Specifically, the average noise rate of BCLCIFAR-10 is 0.23 percentage points, which is comparable to that of ACLCIFAR-10 (0.24 percentage points) and substantially lower than that of the human-annotated CLCIFAR-10 dataset (3.93 percentage points). This result suggests that VLM-based annotation, when combined with a controlled bias design, can achieve favorable noise characteristics while enabling structured complementary-label generation.

Figure 7, 8 and 9 present the histogram of complementary label. The specific imbalance ratio (Cao et al., 2019; Cui et al., 2019) of our dataset can be found in Table 11. This is consistent with the concentrated structure of our transition matrices. Larger label spaces, however, produce a counter-



(a) CIFAR-20 (b) CLCIFAR-20

(c) ACLCIFAR-20 (d) BCLCIFAR-20

Figure 10: Bias-Induced Constrained Labeling transition matrix on CIFAR-20 variants.

Table 11: Imbalance ratios and noise rates for the proposed BICL datasets.

Dataset	**Imbalance Ratio**	**Noise Rate (%)**
BCLCIFAR-10	4,966.5	0.23
BCLCIFAR-20	200.12	0.80
BCLCIFAR-100	119.59	0.026
BCLTinyImageNet-200	416.35	0.006

intuitive effect: a subset of classes receives zero probability in the final complementary-label distribution. This situation could arise from the interaction between our candidate sampling strategy and the inherent biases of VLMs. Specifically, the process involves instructing the VLM with a small subset of candidates alongside a negative prompt. Consequently, low-profile and less semantically distinct classes are excluded from the final dataset.

Crucially, this low-noise characteristic is preserved regardless of the label space size. As summarized in Table 11, our method achieves consistently low error rates across all evaluated datasets.

D ADDITIONAL ABLATION STUDIES

This appendix presents additional ablation studies to evaluate the robustness of BICL across different modeling and annotation choices. We examine the effects of classifier architectures, data augmentation, clustering encoders, VLM annotators, and negative-prompt formulations. We also compare BICL with recent CLL methods, approaches using a small number of true labels, VLM-based true-label collection, and partial-label learning methods. Finally, we analyze how BICL behaves as the number of classes increases to understand its scalability further.

D.1 RESULT ANALYSIS OF DIFFERENT NEURAL NETWORK ARCHITECTURES

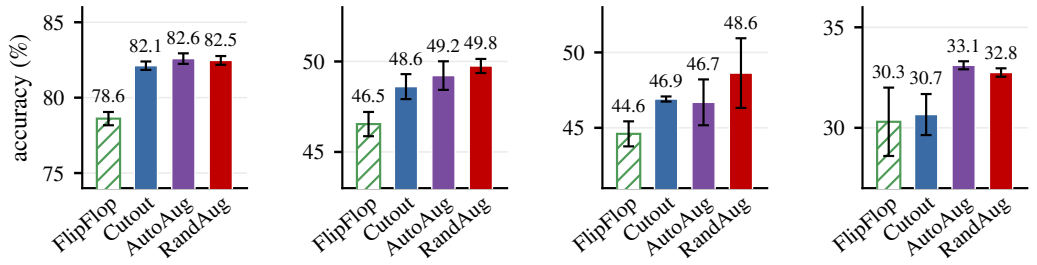
We conducted an ablation study comparing different neural network architectures, including ResNet18, ResNet34, and ResNet50, to determine the most suitable architecture for our complementary dataset. Results shown in Table 12 indicate that ResNet34 achieved the best performance.

Table 12: Performance of various neural network architectures using FWD and CPE-F algorithms on our BICL complementary datasets (accuracy %, mean±std). The best architecture for each dataset–algorithm pair is shown in bold.

Architecture	CIFAR-10		CIFAR-20		CIFAR-100		TinyImageNet-200	
	FWD	CPE-F	FWD	CPE-F	FWD	CPE-F	FWD	CPE-F
ResNet18	78.79±0.91	78.36±0.67	47.68±0.35	46.42±0.54	44.81±0.98	44.36±0.94	31.13±0.33	30.37±0.39
ResNet34	**81.23**±0.05	**80.98**±0.18	**50.84**±0.32	**50.89**±0.30	**46.70**±0.60	**46.57**±0.09	**32.15**±0.30	**31.89**±0.07
ResNet50	78.46±0.20	77.90±0.40	45.68±0.99	44.52±1.20	42.04±0.82	41.59±0.85	30.12±1.07	29.74±0.75

D.2 RESULT ANALYSIS OF DATA AUGMENTATION

We conduct a further analysis on the integration of BICL with various data augmentation techniques to enhance model robustness. Specifically, we evaluate a spectrum of augmentation intensities, ranging from weaker methods, such as Cutout (DeVries & Taylor, 2017), to stronger automated policies like AutoAug (Cubuk et al., 2019) and RandAug (Cubuk et al., 2020). These are compared against the Flipflop baseline, which serves as the standard augmentation strategy employed throughout our experiments in Section 5.



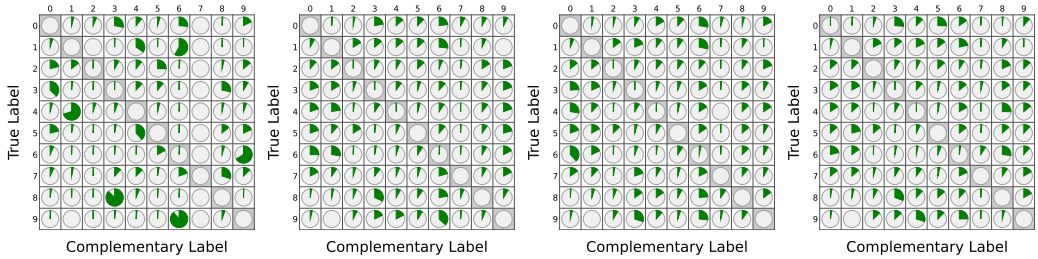
(a) CIFAR-10 (b) CIFAR-20 (c) CIFAR-100 (d) TinyImageNet-200

Figure 11: FWD accuracy of BICL under four data augmentation strategies, from the FlipFlop baseline used throughout Section 5 (green hatched outline) to Cutout (blue), AutoAug (purple) and RandAug (red). Every augmentation improves on FlipFlop on every dataset; the two automated policies give the largest gains, by 2.8–4.0 points.

As illustrated in Figure 11, there is a clear performance improvement of BICL across datasets by incorporating data augmentations compared to the baseline simple Flipflop. Cutout serves as an effective immediate step, performing competitively on easy datasets like CIFAR-10. However, as the complexity of the dataset increases, the gap between Cutout and the stronger augmentation technique widens. RandAug and AutoAug prove to be the most robust out of all strategies tested, constantly delivering a significant boost across all datasets. Overall, the addition of data augmentations to our approach yields an increase in model accuracy regardless of the dataset intricacy.

D.3 Effect of Encoder Choice on BICL Performance

In the CLL setting, ordinary labels are unavailable or costly to obtain. As a result, the clustering induced by the encoder plays a critical role in grouping semantically similar images into collections waiting for VLM annotation.



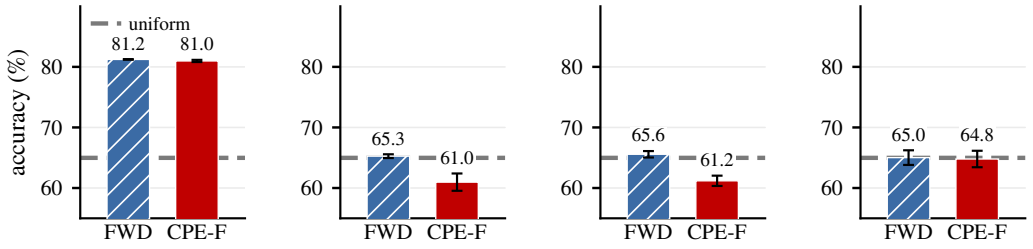
(a) SimSiam (b) SimCLR (c) BYOL (d) MoCov3

Figure 12: Transition matrix induced by the BICL protocol when using different encoder backbones for label selection.

To investigate this, we conducted an ablation study comparing SimSiam (Chen & He, 2021) with other self-supervised encoders, including SimCLR (Chen et al., 2020), BYOL (Richemond et al., 2020) and MoCov3 (Chen et al., 2021). All additional pretraining models were trained with a ResNet18 backbone on CIFAR10 for 800 epochs, using the same hyperparameters as for SimSiam.

Table 13: Information statistics of the CIFAR-10 transition matrices induced by BICL with different self-supervised encoders in the label-selection stage (bits, balanced prior). k is the largest row support and noise the diagonal mass in percent; $H(\bar{Y} \mid Y)$ measures how concentrated the class-conditional outputs are and $\delta = \log C - H(\bar{Y})$ how unevenly the complementary labels are used. All values are computed from the stored matrices with the routine of Appendix A.6.

Encoder	k	noise	$H(\bar{Y} \mid Y)$	δ	$H(Y \mid \bar{Y})$	$I(Y;\bar{Y})$
SimSiam	10	0.23	1.880	0.320	**2.200**	**1.122**
SimCLR	10	0.20	2.818	0.123	2.941	0.381
BYOL	10	0.43	2.828	0.147	2.975	0.347
MoCo v3	10	0.36	2.833	0.095	2.928	0.394



(a) SimSiam (b) SimCLR (c) BYOL (d) MoCov3

Figure 13: BICL performance on CIFAR-10 for each self-supervised encoder used in the label-selection stage, evaluated with FWD (blue, hatched) and CPE-F (red). The dashed grey line marks the uniform-CLL baseline (64.98%), the accuracy reached without BICL. Only SimSiam rises clearly above the baseline, by more than 16 points under both methods; MoCov3 matches it, while SimCLR and BYOL fall roughly 4 points below it under CPE-F.

The efficacy of BICL depends on how well the encoder groups semantically similar images into coherent clusters. Ideally, all samples of a class fall into a single cluster and therefore share one candidate set, which gives a concentrated, class-dependent transition structure. When the encoder

scatters a class across several clusters, that class inherits several candidate sets and its row of the transition matrix becomes diffuse. Figures 12 and 13 show this side by side with the downstream accuracy, and Table 13 quantifies it.

SimSiam induces the most informative transition structure among the evaluated encoders, with the lowest conditional ambiguity and the highest mutual information, and its downstream accuracy follows the same trend. The other three encoders produce far more diffuse class-conditional transition patterns: their rows have the same support size as that of SimSiam ($k = 10$) but are much less concentrated ($H(\bar{Y} \mid Y)$ of 2.82–2.83 bits against 1.88), so they retain less information about the true class. This is consistent with the information-theoretic view of Section 4: what matters is not transition bias as such, but whether the induced mechanism concentrates class-conditional outputs while keeping the complementary labels spread over the label space. The comparison is observational; it shows that encoder quality, transition structure, and accuracy move together rather than establishing a causal chain.

D.4 Effect of VLM Choice on BICL Performance

A core component of our data collection protocol is the leveraging of systematic biases inherent to VLM-based annotation. Different model architectures or pre-training objectives may produce varying degrees of "preference" for certain negative labels, which in turn shapes how concentrated and informative the resulting transition matrix is.

To evaluate the sensitivity of the BICL framework to the underlying VLM, we conduct an ablation study across a diverse suite of contemporary architectures. While our primary experiments utilize LLaVA-v1.6-7B (Liu et al., 2023a), we extend this analysis to include several recent open-weight families, including Qwen3 (Yang et al., 2025) (with various parameter versions) and Gemma4-E4B (Team et al., 2024). Every setting for this extended experiment is kept the same as in Section 5.1, except that the VLM is replaced with alternative models.

Table 14: Accuracy (%) comparison between the uniform baseline and different VLM-based complementary-label generation strategies. Best performance for each dataset–method pair is in bold.

Method	Dataset	Uniform	LLaVA-7B	Qwen2-7B	Qwen3-4B	Qwen3-8B	Gemma4-E4B
FWD	CIFAR-10	64.99±0.64	**81.23**±0.05	79.45±0.82	80.65±0.53	81.00±0.39	78.67±0.72
	CIFAR-20	20.50±0.75	**50.84**±0.32	48.09±0.93	49.58±0.91	51.10±0.41	51.06±0.14
	CIFAR-100	5.53±0.47	**46.70**±0.60	38.26±1.40	36.15±2.69	38.92±0.34	39.13±0.58
	TinyImageNet-200	4.00±0.43	**32.15**±0.30	32.30±1.56	33.50±0.29	34.24±0.25	34.06±1.84
CPE-F	CIFAR-10	64.97±0.61	**80.98**±0.18	79.57±0.86	80.12±0.45	80.68±0.37	77.61±0.69
	CIFAR-20	20.73±1.17	**50.89**±0.30	47.46±1.29	48.36±0.86	49.96±0.25	49.62±0.24
	CIFAR-100	5.06±0.76	**46.57**±0.09	37.81±1.77	36.01±2.73	38.63±0.40	39.04±0.46
	TinyImageNet-200	2.14±0.87	**31.89**±0.07	30.41±1.72	33.37±0.34	33.83±0.13	33.68±2.05

Note: **Uniform** denotes the non-VLM baseline, where CLs are generated under a uniform assumption (Ishida et al., 2017). Our proposed VLM-based settings generally achieve substantially higher accuracy than the uniform baseline.

The results of this ablation study are reported in Table 14. Our designed method consistently outperforms the uniform baseline across all evaluated VLM models, datasets, and learning methods, suggesting that the proposed strategy is robust to different VLM backbones. The model used in our main experiments (LLaVA-v1.6-7B) is the strongest annotator among those evaluated on the CIFAR-based datasets, achieving the highest accuracy in nearly all CIFAR-10, CIFAR-20, and CIFAR-100 runs. This suggests that its visual-linguistic alignment is particularly well-suited for the semantic granularity of these specific label spaces. The picture changes on TinyImageNet-200, where LLaVA performs well but is outperformed by Qwen3-8B and Gemma4-E4B. The newer models appear better able to use the richer visual detail of the higher-resolution TinyImageNet images.

These results indicate that BICL is robust to the choice of VLM and that, on the higher-resolution dataset, it benefits from stronger annotators, which suggests it can improve further as VLMs improve.

D.5 BICL Performance on Additional Negative Prompt

The success of the BICL framework relies on the VLM-based complementary-label annotation phase, where the model is tasked with identifying an incorrect class from a constrained candidate set. As prompting is a critical component of our data collection protocol, we conduct an ablation study to investigate the fragility of prompt engineering and its impact on the resulting downstream classifier performance.

While the base negative prompt used in our main experiments specifically instructs the VLM to "answer the question with a single word" (as detailed in Section 3.1), we observe that even semantically similar prompts can lead to variations in the transition matrix Q, which leads to variations in final classifier performance. For this experiment, we keep all settings identical to those in Section 5.1, but replace the prompt with an alternative formulation in Appendix D.5. Notably, the VLM annotation phase is conducted using LLaVA-v1.6-7B.

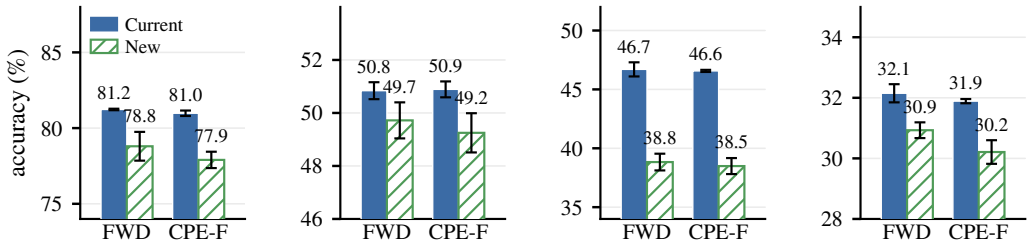
New Prompt

```
Question:<image> Which label does not belong to this image?  (1)
labels[0] (2) labels[1] (3) labels[2] (4) labels[3] Please respond
with only the number of the correct answer.
```

Current Prompt

```
Question:<image> Which label does not belong to this image?  Answer
the question with a single word:  {label[0], label[1], label[2],
label[3]}.
```

Instead of directly naming the complementary label for each image, this prompt requires the VLM to respond with the index corresponding to a given option. From a human perspective, both prompts describe essentially the same task, and we would expect them to yield the same answer for a given image. However, for the VLM, the transition from semantic identification to index-based mapping introduces several layers of inductive and positional bias. Unlike direct naming, which leverages the strong visual-textual alignment of the VLM, the multiple-choice format requires an additional step that maps a semantic class to an arbitrary numerical index.



(a) CIFAR-10 (b) CIFAR-20 (c) CIFAR-100 (d) TinyImageNet-200

Figure 14: Accuracy of BICL under the current prompt used in our experiments (blue) and the alternative index-based prompt (green hatched outline), for FWD and CPE-F on each dataset. Each panel has its own accuracy range. The current prompt is better in every setting, by 1–3 points on CIFAR-10, CIFAR-20 and TinyImageNet-200 and by about 8 points on CIFAR-100.

The choice of negative prompt strongly influences the supervision the VLM produces. As Table 15 and Figure 14 show, two formulations of the same labeling task can lead to clearly different downstream accuracy. The largest effect is on CIFAR-100, where the alternative prompt yields a matrix with about half the mutual information of the current one (1.05 against 2.07 bits) and the accuracy drops by about 8 percentage points. On CIFAR-10 the alternative prompt also loses information and accuracy, whereas on CIFAR-20 and TinyImageNet-200 the two prompts induce nearly the same information statistics and the accuracy gap is small. Within this ablation, lower mutual information therefore goes together with lower accuracy, in line with the view of Section 4, although that analysis

Table 15: Label noise and information statistics of the BICL matrices collected with the current prompt and with the alternative prompt (bits, balanced prior). The first row is the diagonal mass of the stored matrix in percent, i.e., the label-noise rate under a balanced prior. The better value of each pair is in bold.

Datasets	**CIFAR-10**		**CIFAR-20**		**CIFAR-100**		**TinyImageNet-200**	
Prompt	Current	Alternative	Current	Alternative	Current	Alternative	Current	Alternative
Diagonal mass of Q (%)	**0.23**	0.64	**0.80**	1.03	**0.026**	0.14	0.055	**0.036**
$H(\bar{Y} \mid Y)$	**1.880**	2.067	**2.884**	3.175	**4.100**	5.097	**4.453**	4.553
$H(Y \mid \bar{Y})$	**2.200**	2.460	3.618	**3.602**	**4.573**	5.590	**5.839**	5.868
$I(Y;\bar{Y})$	**1.122**	0.862	0.704	**0.720**	**2.071**	1.054	**1.805**	1.776

does not by itself predict accuracy. These findings underscore the fragility of prompt engineering: the quality of the supervision signal is sensitive to the linguistic framing of the task.

D.6 BICL Performance on Recent CLL Methods

In addition to the standard CLL algorithms for biased design, we conduct additional experiments with recent CLL baselines, including SCARCE (Wang et al., 2024a) and OP (Liu et al., 2023b). We want to emphasize that these methods belong to two different technical directions in the CLL literature. Methods such as SCL-NL, SCARCE, COVR (You et al., 2025), and OP are primarily developed for the standard CLL setting, where the complementary-label transition is assumed to be relatively uniform or unbiased. In contrast, methods such as FWD, CPE, URE-TNN, and URE-TGA are designed to model or handle biased transition structures.

Since BICL induces a strongly non-uniform complementary-label distribution, we consider the second group to be more appropriate for the main comparison. For completeness, we also evaluated recent methods such as SCARCE and OP, although their assumptions are less well aligned with the setting induced by BICL, as shown in Table 16.

Table 16: Performance comparison of recent CLL methods under uniform and BICL distributions on CIFAR-20 and CIFAR-100 (accuracy (%), mean±std).

Datasets	**CIFAR-20**		**CIFAR-100**	
Method	Uniform (%)	BICL (%)	Uniform (%)	BICL (%)
SCARCE	17.53±0.49	5.02±0.10	1.41±0.07	1.01±0.02
OP	20.71±1.04	5.48±0.43	1.21±0.18	1.02±0.05
FWD	20.50±0.75	50.84±0.32	5.53±0.47	46.70±0.60
CPE-F	20.73±1.17	50.89±0.30	5.06±0.76	46.57±0.09

Table 16 shows that recent CLL methods such as SCARCE and OP perform poorly under the BICL distribution, suggesting limited robustness to strongly non-uniform complementary-label collection. On CIFAR-20, their accuracy drops substantially from 17.53% and 20.71% under the uniform setting to 5.02% and 5.48%, respectively. A similar trend is observed on CIFAR-100, where both methods remain near chance level under BICL. In contrast, FWD and CPE-F benefit markedly from the BICL distribution, achieving large improvements over the uniform setting on both datasets. These results indicate that strongly non-uniform complementary-label distributions can severely hinder methods built for the uniform setting. At the same time, our setting remains effective for more robust baselines such as FWD and CPE-F.

D.7 Comparison with Approaches Using Few True Labels

In CLL, true labels are not accessible during training. However, several *transition-aware* (Q-aware) algorithms (Yu et al., 2018; Lin & Lin, 2023) assume that the complementary-label transition matrix Q is known or can be reliably estimated. In practice, this assumption is often violated. When learning solely from CLs, information about the true label distribution is largely lost, making it

fundamentally difficult to estimate Q without additional supervision, as Q is closely tied to the class priors of true labels.

On the other hand, while annotating large-scale datasets with full supervision is costly, it is often feasible in practice to obtain a small number of confidently labeled examples per class. Motivated by this observation, we consider a setting in which five true labels per class are manually provided and used to estimate the transition matrix. In this sense, Q-aware CLL methods implicitly rely on a small amount of true-label supervision.

To ensure a fair comparison, we conduct an ablation study in which five true labels per class are used to train a classifier under standard supervision, while the remaining data are treated as complementarily labeled. Both the true-label and complementary-label distributions are generated uniformly and without noise. Following Ishida et al. (2017), we combine the losses from ordinary labels and CLs as

$$\hat{R}(f) = \underbrace{\frac{\alpha}{m}\sum_{j=1}^{m}\mathcal{L}\big(f(\mathbf{x}_j), y_j\big)}_{CE} + \underbrace{\frac{(1-\alpha)(K-1)}{n}\sum_{i=1}^{n}\mathcal{L}\left(Q^{\top}\sigma(g(\mathbf{x}_i)), \bar{y}_i\right)}_{FWD},$$

where $\{(\mathbf{x}_j, y_j)\}_{j=1}^{m}$ denotes the ordinarily labeled data and $\{(\mathbf{x}_i, \bar{y}_i)\}_{i=1}^{n}$ denotes the complementarily labeled data, $\alpha \in [0, 1]$ is a hyper-parameter that interpolates between the two risks; we set $\alpha = 0.5$ in this experiment. Q is the estimated transition matrix, and $g(\mathbf{x})$ denotes the classifier outputs (logits) and $\sigma(\cdot)$ is the softmax function; $\mathcal{L}(\cdot, \cdot)$ is a standard loss function (i.e., Cross-Entropy).

Table 17: Performance comparison with methods using the small seed set of true labels (5 per class) combined with uniform CLs (CE + FWD) against BICL (FWD).

True Labels	Algorithm	CIFAR-10	CIFAR-20	CIFAR-100	TinyImageNet-200
5 (Seed Set)	CE + FWD	26.94	11.28	8.33	3.62
BICL (Ours)	FWD	**81.23**	**50.84**	**46.70**	**32.15**
	Δ	$\uparrow$ 54.29	$\uparrow$ 39.56	$\uparrow$ 38.37	$\uparrow$ 28.53

CE+FWD: In this setup, we train on the small set of true-labeled samples (five per class) using cross-entropy (CE), and on the remaining samples using FWD with uniform distribution (Ishida et al., 2017). The final objective is the sum of the CE and FWD losses. In contrast, FWD in BICL (ours) is trained on biased CLs collected using our BICL protocol.

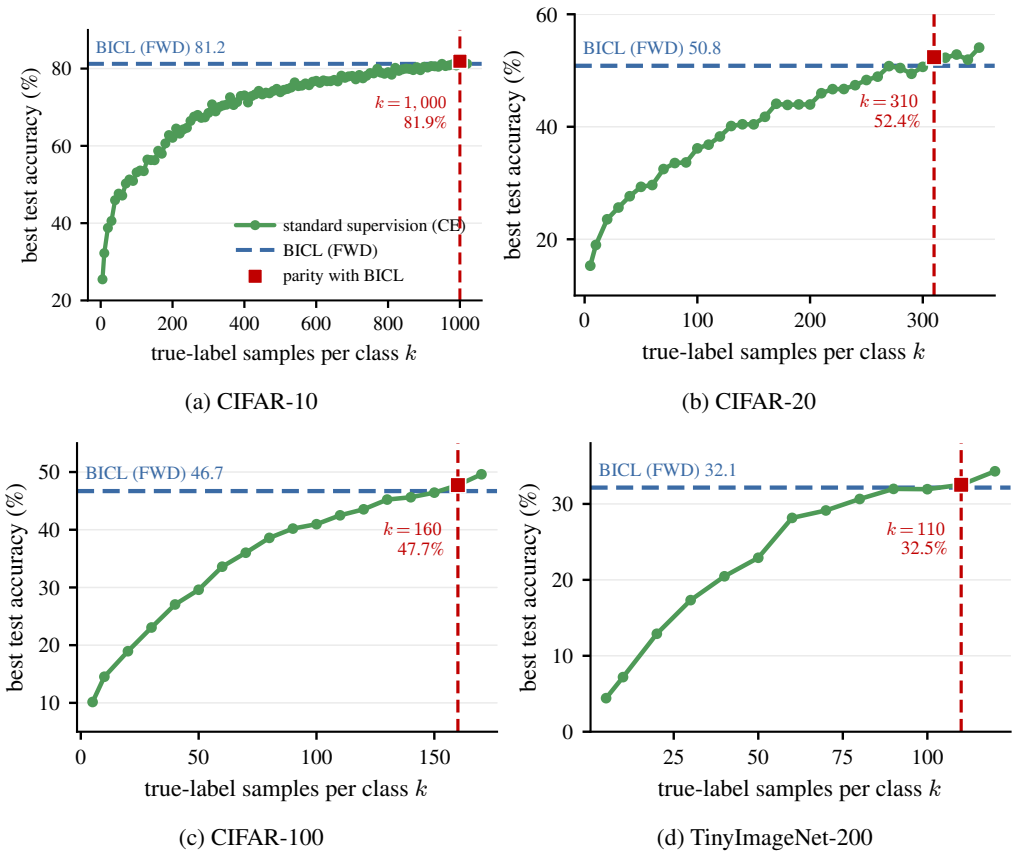
We use ResNet34 as the backbone network and adopt the same hyperparameter settings described in Section 5.1. The results in Table 17 show that BICL consistently outperforms the combined-loss baseline (CE+FWD) that leverages five true labels per class across all datasets. These findings indicate that our approach is more effective than relying on a small amount of true-label supervision combined with massive CLs, highlighting the advantage of BICL in settings where limited true labels are available.

D.8 True-Label Budget Required for Standard Supervision to Reach BICL Accuracy

While complementary labels only provide indirect negative evidence about the true class, our results in Section 5 show that BICL achieves strong classification performance across diverse datasets. A natural practical question arises: *How many true-labeled examples does standard supervised cross-entropy training require to reach the accuracy obtained by BICL*?

To quantify this budget, we train a classifier under standard supervised learning using the standard CE loss on varying quantities of true labeled samples. For each dataset, we vary the number of true-labeled training instances per class, k, from small few-shot budgets ($k \geq 5$) up to the point where ordinary classification accuracy matches that of BICL trained with the FWD algorithm. All training hyperparameters, data augmentations, and network backbones (ResNet-34) are kept identical to the baseline configuration in Section 5.1.

As shown in Table 18 and Figure 15, standard supervised learning with CE requires substantially more true labels to reach accuracy comparable to BICL across all four benchmarks. Specifically,



(a) CIFAR-10 (b) CIFAR-20

(c) CIFAR-100 (d) TinyImageNet-200

Figure 15: Sample efficiency of standard supervision against BICL. Each panel plots the best test accuracy of a ResNet-34 trained with cross-entropy on k true label samples per class (green, one run per k) against the accuracy BICL reaches with FWD on the full complementary-label dataset (blue dashed). The red square and dashed drop line mark parity, the smallest tested k at which standard supervision matches BICL: 1,000, 310, 160 and 110 samples per class, i.e. 20%, 12.4%, 32% and 22% of each training set with true labels.

Table 18: Number of true labels per class required under standard supervised learning to match or surpass BICL accuracy across four benchmark datasets.

Dataset	**BICL (%)**	**CE (%)**	**True Labels / Class**	**Total Dataset Ratio (%)**
CIFAR-10	81.23±0.05	81.90	1,000	20.0%
CIFAR-20	50.84±0.32	52.38	310	12.4%
CIFAR-100	46.70±0.60	47.74	160	32%
TinyImageNet-200	32.15±0.30	32.52	110	22%

matching BICL requires true labels for 12.4% to 32.0% of the training data, corresponding to 16,000 images on CIFAR-100 and 22,000 on TinyImageNet-200. Standard supervised learning exhibits strong initial gains that gradually diminish as the number of true labels increases, while low-budget CE (e.g., $\leq$ 50 samples/class) remains well below BICL performance. Standard CE trained only on true-labeled examples thus requires substantially larger true-label budgets to reach the accuracy obtained by BICL in our setting. This comparison measures the true-label annotation budget rather than information equivalence, since BICL additionally exploits complementary labels for the remaining training pool.

D.9 Compared to True Label Collection Approach with VLM

To better quantify the gap between collecting *true labels* and *complementary labels* with a VLM, we conduct an ablation study in which we replace complementary-label collection with true-label collection using the same VLM. We then train a classifier under *standard supervision* using the cross-entropy loss and compare its performance against CLL under our BICL protocol. Across datasets, BICL outperforms this standard-supervision baseline, where the model is trained directly on VLM-collected true labels. We attribute this gap to the fact that, in large label spaces, VLMs often struggle to identify the correct class among many visually similar alternatives precisely. In contrast, BICL simplifies the decision process of the VLM by framing annotation as a constrained *rejection* task and restricting the candidate label space through the proposed data collection protocol.

This advantage yields large gains on challenging multi-class benchmarks: compared with standard supervision, BICL improves accuracy by 17.65 percentage points on CIFAR-20, 10.11 percentage points on CIFAR-100, and nearly doubles accuracy on TinyImageNet-200 (see the *Std-Supervision* row in Tables 3 and 4). Overall, these results suggest that in scalable settings, determining what an image is *not* can be more reliable for VLMs than determining what it *is*.

D.10 Partial-Label Learning (PLL) vs. CLL under BICL

While CLL is technically a subclass of PLL, it is the most difficult problem of PLL. To examine this connection empirically, we conducted additional experiments with representative PLL methods, including PiCO (Wang et al., 2022b), PRODEN (Lv et al., 2020), and SoLar (Wang et al., 2022a), on the BICL datasets.

In these experiments, we converted every single complementary label provided by the VLM into a partial-label set by treating all remaining labels in the label space as candidate labels. Thus, each instance in BCLCIFAR-20 is associated with a partial-label set containing 19 candidate labels, and each instance in BCLCIFAR-100 is associated with a partial-label set containing 99 candidate labels. We trained each PLL method for 800 epochs with learning rate 10^{-3}, weight decay 10^{-4}, and momentum 0.9. The corresponding results are reported in Table 19.

Table 19: Comparison of PLL and CLL methods under BICL on BCLCIFAR-20 and BCLCIFAR-100.

Setting	Method	Dataset	Label Set Size	Accuracy (%)
PLL Algorithms	PiCO	BCLCIFAR-20	19	6.01
	PRODEN	BCLCIFAR-20	19	7.74
	SoLar	BCLCIFAR-20	19	4.74
	PiCO	BCLCIFAR-100	99	1.15
	PRODEN	BCLCIFAR-100	99	1.01
	SoLar	BCLCIFAR-100	99	1.10
CLL Algorithms	FWD	BCLCIFAR-20	1	50.84
	CPE-F	BCLCIFAR-20	1	50.89
	FWD	BCLCIFAR-100	1	46.70
	CPE-F	BCLCIFAR-100	1	46.57

The results show that PLL methods perform very poorly in this setting. On BCLCIFAR-20, PiCO, PRODEN, and SoLar achieve 6.01%, 7.74%, and 4.74% accuracy, respectively, which is far below the performance of the CLL method FWD (50.84%). On BCLCIFAR-100, PLL performance further drops to around 1%, while FWD still achieves 46.70%.

This issue becomes even more severe as the number of classes increases. For example, the conversion yields candidate label sets of size 99 for CIFAR-100 and 199 for TinyImageNet-200, making the supervision increasingly ambiguous. Thus, although multiple complementary-label learning is more closely related to PLL, the single-complementary-label setting considered here is much more

difficult for PLL methods after conversion, because the candidate set contains almost the entire label space.

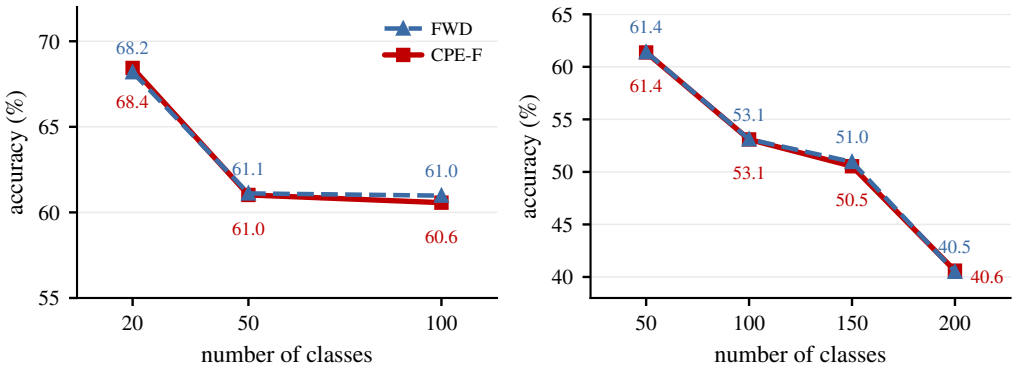
In addition, most PLL methods assume that candidate labels are generated in a relatively random or unbiased manner. By contrast, BICL induces a strongly non-uniform candidate-label structure. Even with large candidate sets, PLL methods can still succeed when the inclusion pattern provides a reliable signal that helps distinguish the true label from incorrect ones. That signal becomes much weaker in our setting due to the strong bias introduced by BICL, which further explains the poor performance of PLL methods.

More generally, this highlights an interesting research direction connecting CLL and PLL. To the best of our knowledge, most PLL methods have not been systematically evaluated in large-scale settings with very large candidate label sets.

D.11 Effect of the Number of Classes on Performance

In CLL, the number of classes plays an important role, as a larger label space generally makes the learning problem more challenging. To better understand how performance degrades as a function of label-space size under CLL, we conduct additional ablation studies to examine further the impact of class count within a fixed dataset. Specifically, we construct two subsets from CIFAR-100, comprising 20 and 50 classes, respectively. The subset preserves the original images and their associated labels, while restricting the overall label space to a smaller set of classes. The 20-class subset of CIFAR-100 is constructed by selecting one randomly chosen fine-grained class from each coarse class (*i.e., the CIFAR-20 superclass grouping*). For the 50-class subset, all classes are randomly selected. To further investigate the scalability of BICL on even larger label spaces, we introduce three additional subsets derived from TinyImageNet-200, containing 50, 100, and 150 classes, respectively. Similar to the 50-class CIFAR-100 configuration, all classes in these TinyImageNet subsets are randomly selected from the original global label space.

In this experiment, the VLM is asked to provide a complementary label by selecting from four options, under the assumption that the *true label* is not included (as in the BICL Analysis Case). All other parameters were kept the same as in our previous experiments. As shown in Figure 16, the results confirm that model performance consistently decreases as the number of classes increases.



(a) CIFAR-100 (b) TinyImageNet-200

Figure 16: Effect of label-space size on performance within CIFAR-100 and TinyImageNet-200. Each panel varies the number of classes inside a single dataset, evaluated with FWD (blue, dashed, triangles) and CPE-F (red, solid, squares). Accuracy falls as the label space grows on both datasets, and the two methods trace the same profile, differing by at most 0.41 points. CIFAR-100 loses 7.5 points between 20 and 100 classes, whereas TinyImageNet-200 loses 20.8 points between 50 and 200.

The most prominent trend across all four figures is the clear inverse correlation between label-space cardinality and classification accuracy. On both CIFAR-100 and TinyImageNet-200 subsets, we observe that accuracy drops markedly as the number of classes increases, reflecting the escalating difficulty of the classification task. Notably, the FWD and CPE-F algorithms exhibit nearly iden-

tical performance profiles across all subsets, suggesting that these trends are driven by inherent data complexity. In conclusion, while the BICL framework successfully enables CLL to scale to high-cardinality environments, the mounting supervision ambiguity in larger label spaces remains a critical bottleneck for performance stability.

E Details of Experimental Setup

This section describes the experimental configuration used to evaluate the proposed method. We first detail the VLMs employed for automated annotation and the benchmark datasets used in our experiments. We then summarize the compared approaches and CLL algorithms, followed by the architectural details of the models used for training. Together, these descriptions ensure reproducibility and provide sufficient context for interpreting the empirical results.

E.1 Details of Vision–Language Models

For automated annotation, we use LLaVA-v1.6-7B (Liu et al., 2023a) (Large Language-and-Vision Assistant). We choose LLaVA to enable a direct comparison with ACLImage (Mai et al., 2025), which also adopts this model for complementary-label collection. In particular, ACLImage reports that LLaVA offers a favorable trade-off between computational cost and labeling quality for CLL. From a technical perspective, the LLaVA architecture bridges a pre-trained vision encoder (such as CLIP ViT-L/14) and a large language model via a simple linear or MLP projection matrix. We use the 7B-parameter variant of version 1.6, which introduces dynamic high-resolution image processing and enhanced visual reasoning, providing an optimal balance between vision reasoning capability and resource efficiency.

To further evaluate the sensitivity and robustness of our framework across diverse multimodal architectures, our extended ablation studies incorporate recent open-weight models, specifically from the Qwen3 (Yang et al., 2025) and Gemma (Team et al., 2024) families. The Qwen3 architecture distinguishes itself through its native dynamic-resolution visual encoding and sophisticated visual token compression mechanism. Rather than forcing images into a fixed grid, this standout methodology allows Qwen3 to optimally preserve aspect ratios and fine-grained details at massive resolutions without overwhelming the context window. Similarly, we evaluate Gemma-based vision models (e.g., the Gemma4-E4B variant). The defining hallmark of this Gemma lineage is its deep, seamless integration of the highly optimized SigLIP (Sigmoid Loss for Language Image Pre-training) vision encoder. This architectural choice yields extreme parameter efficiency, allowing a compact model to achieve the granular spatial awareness and dense visual reasoning typically reserved for much larger architectures. Incorporating these modern models demonstrates that our data collection protocol scales effectively as underlying annotators gain specialized, cutting-edge abilities to extract rich features and process high-definition images.

All VLM queries are performed in a zero-shot setting with temperature set to 0. The VLM receives each input image at its original resolution (as downloaded). We instruct the model using a task-specific prompt derived from the base *negative prompt* in Section 3.1, with minor modifications depending on the experiment. For Figure 6, we use the same base prompt while expanding the candidate label options to match the required number of labels n.

In the standard supervision experiments, the VLM acts as a zero-shot classifier over the global label space. The model is explicitly instructed to find the true label using a positive prompt in Appendix E.1. Since all the images were raw, LLaVA actually struggles to answer this task consistently and frequently returns reasons such as "*This image is too blurry, I can not select the true label*". After rigorous prompting, the final instruction enforced a forced-choice mechanism in the selection phase, eliminating uncertain outputs.

Prompt

Question:<image> Select the closest label from [labels]. Output strictly one label, no explanation or reasoning. If the image is unclear, choose the most probable label.

E.2 Details of Benchmark Datasets

In this section, we report the index-to-class-name mappings for the benchmark datasets used throughout our experiments. Table 20 lists the class indices for CIFAR-10 and the corresponding superclass indices for CIFAR-20, following the standard CIFAR-100 taxonomy where CIFAR-20 groups the CIFAR-100 fine classes into 20 superclasses. Table 21 provides the complete mapping between class indices (0–99) and fine-grained label names for CIFAR-100. These mappings are included to ensure clarity and reproducibility when interpreting results reported by class index (i.e., in error analysis and ablation studies) and to facilitate direct comparison across datasets with different label granularities.

Table 20: The correspondence between index and label names of CIFAR-10 and CIFAR-20 datasets.

Index	CIFAR-10 Label Name	CIFAR-20 Label Name
0	airplane	aquatic mammals
1	automobile	fish
2	bird	flowers
3	cat	food containers
4	deer	fruit, vegetables and mushrooms
5	dog	household electrical devices
6	frog	household furniture
7	horse	insects
8	ship	large carnivores and bear
9	truck	large man-made outdoor things
10		large natural outdoor scenes
11		large omnivores and herbivores
12		medium-sized mammals
13		non-insect invertebrates
14		people
15		reptiles
16		small mammals
17		trees
18		transportation vehicles
19		non-transportation vehicles

E.3 Descriptions of Compared Approaches

In this subsection, we provide detailed descriptions of the existing real-world complementary-label datasets used for comparison: CLImage (Wang et al., 2025) and ACLImage (Mai et al., 2025). We further discuss the methodological differences regarding the specific annotators (human vs. VLM) used in these approaches compared to our proposed BICL.

E.3.1 CLImage: Human-Annotated Complementary Labels

CLImage (Wang et al., 2025) represents the first effort to collect CLs from human annotators to capture real-world noise and bias distributions. The dataset collection process involves utilizing crowdsourcing platforms (i.e., Amazon Mechanical Turk) where human workers are presented with images and asked to select a label that does *not* match the image content.

Protocol. Annotators are presented with a randomly sampled subset of candidate labels (specifically 4 labels) from the label space and asked to select one incorrect class. This process relies on human subjectivity, where annotators may select CLs based on visual dissimilarity or random choice.

Characteristics. According to recent benchmarks (Ye et al., 2024), CLImage exhibits a relatively high label noise rate (i.e., 3.93 percentage points on CIFAR-10) due to human error. However, interestingly, it maintains a relatively low imbalance ratio (approx. 1.56), suggesting that human errors and choices are somewhat more evenly distributed across classes compared to model-based annotation.

Table 21: Fine-grained class indices and names of the CIFAR-100 dataset.

Index	Label Name	Index	Label Name	Index	Label Name	Index	Label Name
0	apple	25	couch	50	mouse	75	skunk
1	aquarium_fish	26	crab	51	mushroom	76	skyscraper
2	baby	27	crocodile	52	oak_tree	77	snail
3	bear	28	cup	53	orange	78	snake
4	beaver	29	dinosaur	54	orchid	79	spider
5	bed	30	dolphin	55	otter	80	squirrel
6	bee	31	elephant	56	palm_tree	81	streetcar
7	beetle	32	flatfish	57	pear	82	sunflower
8	bicycle	33	forest	58	pickup_truck	83	sweet_pepper
9	bottle	34	fox	59	pine_tree	84	table
10	bowl	35	girl	60	plain	85	tank
11	boy	36	hamster	61	plate	86	telephone
12	bridge	37	house	62	poppy	87	television
13	bus	38	kangaroo	63	porcupine	88	tiger
14	butterfly	39	keyboard	64	possum	89	tractor
15	camel	40	lamp	65	rabbit	90	train
16	can	41	lawn_mower	66	raccoon	91	trout
17	castle	42	leopard	67	ray	92	tulip
18	caterpillar	43	lion	68	road	93	turtle
19	cattle	44	lizard	69	rocket	94	wardrobe
20	chair	45	lobster	70	rose	95	whale
21	chimpanzee	46	man	71	sea	96	willow_tree
22	clock	47	maple_tree	72	seal	97	wolf
23	cloud	48	motorcycle	73	shark	98	woman
24	cockroach	49	mountain	74	shrew	99	worm

E.3.2 ACLIMAGE: VLM-ANNOTATED COMPLEMENTARY LABELS

To address the scalability issues and high costs of human annotation, ACLImage (Mai et al., 2025) automates the collection process using VLMs, specifically leveraging models like **LLaVA** (Liu et al., 2023a).

Protocol. ACLImage adopts the same protocol as CLImage to generate CLs. For each image:

1. A candidate set of 4 labels is uniformly sampled from the label space.
2. The VLM (LLaVA) is prompted to select the label from this set that does *not* describe the image.

Characteristics. While using the same protocol, the VLM annotator substantially reduces label noise, achieving a noise rate of approximately 0.24 percentage points on CIFAR-10, which is much lower than CLImage. However, contrary to the expectation that uniform candidate sampling would yield a uniform distribution, ACLImage exhibits a *higher* imbalance ratio (approx. 3.26) compared to CLImage (1.56). This indicates that even with restricted candidates, VLMs exhibit inherent biases towards selecting specific "safe" or common negative labels.

E.3.3 COMPARISON WITH BICL

Our proposed BICL differs fundamentally from these approaches in design intent and resulting distribution.

Bias Structure. While ACLImage exhibits moderate bias (imbalance ratio ≈ 3.26) as an unintended side effect of VLM preferences, BICL deliberately *induces* a strongly concentrated structure (imbalance ratio ≈ 4966.5) through its cluster-based candidate selection. This structured signal carries

more information about the true class than the CLImage or ACLImage distributions (Table 7: 1.12 bits per label on CIFAR-10 against 0.08 and 0.30).

Noise vs. Bias Trade-off. CLImage represents a high-noise, low-bias regime (Human). ACLImage represents a low-noise, moderate-bias regime (LLaVA). BICL represents a low-noise, high-bias regime (LLaVA + Clustering). In our experiments, the low-noise, strongly structured regime of BICL yields the highest accuracy with transition-aware learners; Section 4 makes clear that it is the structure of the transitions, not the amount of bias as such, that carries the information.

E.4 Details of Model Architecture

For all experiments, we employed ResNet architectures (ResNet-18, ResNet-34, and ResNet-50) as the backbone networks. To accommodate the lower resolution of the CIFAR (32×32) and TinyImageNet (64×64) datasets compared to standard ImageNet inputs, we modified the initial stem of the network. Specifically, we replaced the original 7×7 convolutional layer (stride 2) with a 3×3 convolutional layer with stride 1 and padding 1. This modification prevents the loss of fine-grained spatial information during the early downsampling stages. The network weights were initialized using Kaiming initialization (He et al., 2016).

E.5 Complementary-Label Algorithms

In this section, we review the six CLL algorithms used as baselines in our evaluation. These methods differ in whether and how they exploit the transition matrix Q relating latent true labels to observed CLs.

E.5.1 Forward Correction (FWD).

Proposed by Yu et al. (2018), the Forward Correction (FWD) method provides a way to learn from CLs by correcting the loss function using the transition matrix Q. Instead of predicting the complementary label directly, FWD utilizes the transition matrix to map the predicted probability of the true label to the probability of the complementary label.

Let $g(\mathbf{x})$ denote the classifier outputs (logits) and $\sigma(\cdot)$ be the softmax function. The loss function for FWD is defined as:

$$R_{\text{FWD}}(g) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(Q^\top \sigma(g(\mathbf{x}_i)), \bar{y}_i\right),$$

where N is the number of samples, $\bar{y}_i$ is the complementary label, and $\mathcal{L}(\cdot, \cdot)$ is a standard loss function (i.e., Cross-Entropy). By incorporating Q, the model learns to approximate the true class distribution even when supervision is provided only via CLs.

E.5.2 Unbiased Risk Estimator (URE).

Ishida et al. (2019) proposed a framework to estimate the classification risk unbiasedly using CLs. The core idea relies on an inverse transition matrix to recover the unbiased risk of the true classifier. The general loss formulation is:

$$R_{\text{URE}}(g) = \frac{1}{N} \sum_{i=1}^{N} e_{\bar{y}_i}^\top Q^{-1} \mathcal{L}(g(\mathbf{x}_i)),$$

where $e_{\bar{y}_i}$ is the one-hot vector of the complementary label, and $\mathcal{L}(g(\mathbf{x}_i))$ denotes the vector of losses for all possible classes. Because this formulation explicitly involves Q^{-1}, estimation errors may be amplified when Q is poorly conditioned, potentially increasing the variance of the risk estimator. However, this unbiased estimator can result in negative empirical risk values during training, leading to overfitting. To address this, two correction mechanisms were introduced:

URE-TNN (Non-Negative) The Non-Negative (NN) correction imposes a constraint that the estimated risk for any class cannot be negative. If the risk estimate for a specific class becomes negative during a training iteration, it is clipped to zero to maintain stability.

URE-TGA (Gradient Ascent) The Gradient Ascent (GA) correction strategy takes a different approach. When the empirical risk for a specific class is negative, the algorithm performs gradient ascent (reversing the gradient direction) on that component. This prevents the model from becoming over-confident in incorrect predictions caused by the negative risk estimation.

E.5.3 COMPLEMENTARY PROBABILITY ESTIMATION (CPE).

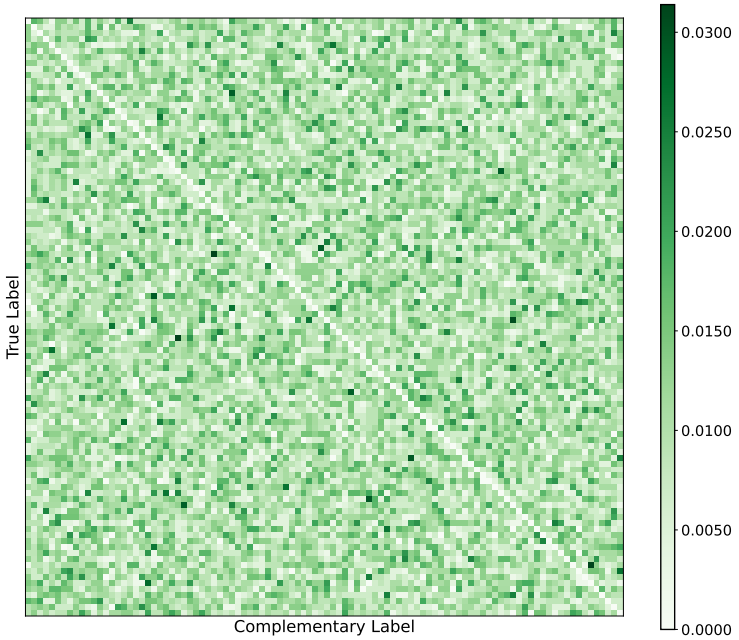
Lin & Lin (2023) introduced the Complementary Probability Estimation (CPE) framework. Unlike risk-correction methods, CPE focuses on directly estimating the probability of a label being complementary, denoted as $p(\bar{y}|\mathbf{x})$. The objective is to minimize the divergence between the model output and the complementary target. CPE employs a surrogate complementary estimation loss (SCEL) as:

$$R_{\text{CPE}}(\mathbf{g}) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(\mathbf{g}(\mathbf{x}_i), e_{\bar{y}}).$$

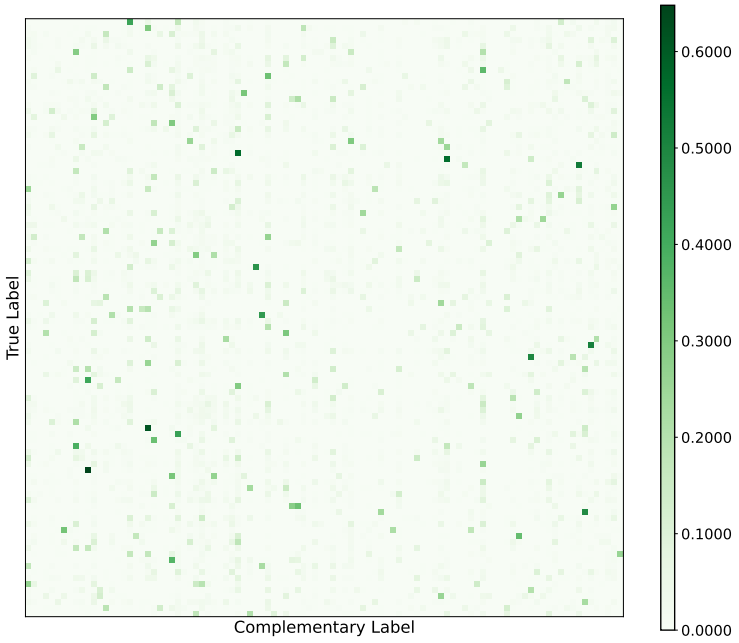
CPE-I (Identity) CPE-I utilizes a direct mapping where no explicit transition layer is applied to the model output before loss calculation. It treats the complementary probability estimation directly without structural modification involving the transition matrix Q.

CPE-F (Fixed) CPE-F incorporates a fixed transition layer derived from the known or estimated transition matrix Q. The model predicts the ordinary label distribution, which is then passed through this fixed layer to produce complementary probabilities $p(\bar{y}|\mathbf{x})$ for training.

CPE-T (Trainable) CPE-T extends CPE-F by making the transition layer trainable. This allows the model to jointly learn the classifier and fine-tune the transition matrix simultaneously, providing robustness against noise or inaccuracies in the initial transition matrix assumption.



(a) CIFAR-100



(b) BCLCIFAR-100

Figure 17: Bias-Induced Constrained Labeling transition matrix on CIFAR-100 variants.