The Unexplored Potential of Vision-Language Models for Generating Large-Scale Complementary-Label Learning Data

Tan-Ha Mai*, Nai-Xuan Ye*, Yu-Wei Kuan, Po-Yi Lu, and Hsuan-Tien Lin(✉)

National Taiwan University, Taipei, Taiwan
{d10922024, b09902008, b11902132, d09944015, htlin}@csie.ntu.edu.tw

Abstract. Complementary-Label Learning (CLL) is a weakly-supervised learning paradigm designed to reduce label collection costs compared to traditional supervised learning with ordinary labels. However, its competitiveness and feasibility in real-world scenarios still need to be determined. Although recent CLL studies using real-world datasets with human annotations have begun to explore these challenges, annotating complementary labels still incurs a non-trivial cost. Consequently, the current availability of real-world data is insufficient to fully demonstrate the practical scalability of CLL. The emergence of Vision-Language Models (VLMs) presents a promising alternative to address the limitation. Somehow, our analysis shows that directly converting the human labeling process for VLMs introduces significant label noise and bias. To address this issue, we developed customized prompts designed to systematically reduce label noise and bias in VLM-based labeling. Our proposed framework effectively curates VLM-annotated, achieving an improvement of 10% performance over human-annotated datasets. This work represents a significant step toward making CLL viable for real-world applications.

Keywords: VLM annotation · Complementary Datasets · Complementary-Label Learning.

1 Introduction

Complementary-Label Learning (CLL) is a paradigm designed to address the high costs of acquiring ordinary labels, a major challenge in multi-class classification applications. Obtaining ordinary labels can be prohibitively expensive, time-consuming, and reliant on expert annotators in some applications. In contrast, complementary labels (CLs)–annotations that indicate only the categories a data point does not belong to [1,2]–can be collected with significantly lower cost and effort. This potential has inspired extensive research into learning from CLs, leading to the development of algorithms grounded in theoretical frameworks. Many studies have demonstrated that models can effectively learn from complementary labels alone, achieving promising results on synthetic datasets [3–6].

* Equal contribution.

Despite significant algorithms and theoretical advancements, early research mainly relied on synthetic datasets. The exclusive use of synthetic datasets in initial studies left the practical effectiveness of CLL methods largely untested in real-world scenarios, raising critical concerns about their applicability. To address this gap, researchers have shifted their focus to real-world scenarios, resulting in the development of CLImage [7], the first set of human-annotated complementary-label collection of datasets designed to reflect real-world distributions. CLImage offered an in-depth analysis of the characteristics of human-annotated CLs and evaluated the performance of existing CLL methods on these datasets. Their findings highlighted that inherent biases—such as human annotators favoring easily recognizable items—and label noise, where ordinary labels appear as complementary labels, can substantially impair the performance of existing algorithms [7].

While CLL research has predominantly focused on image datasets, it has yet to extend to other modalities, such as text or video. The emergence of Vision-Language Models (VLMs) [8, 9] offers a promising alternative to human annotation. In contrast to Large Language Models (LLMs) [10], which are optimized for text-based tasks, VLMs are specifically designed to process multimodal data, making them particularly suited for complementary-label annotation in vision-centric domains. Recent studies have demonstrated the utility of VLMs across applications, including multi-label learning [11], semi-supervised learning [12], and learning from partial labels [13, 14], showcasing their potential to address the limitations of current CLL methods effectively.

To the best of our knowledge, no prior work has systematically investigated *how to effectively adapt and utilize VLMs for annotating weak labels.* This gap in the literature serves as the foundation for our study, which aims to propose and evaluate a novel framework for leveraging VLMs to annotate complementary-label datasets. Our analysis revealed that a direct adaptation of the human-labeling protocol for VLM-based annotation encounters a challenge, particularly with high label noise rates. Label noise is a key factor that degrades the performance of learning classifiers in CLL [7, 15]. This suggests that human-labeling protocols are inefficient when directly applied to VLM-based annotation. In response, we developed a tailored complementary label collection protocol specifically optimized for VLMs. Our proposed protocol achieved remarkable success, reducing label noise rates compared to the human-annotated CLImage datasets [7]. This reduction in label noise highlights the potential of our proposed method to enhance the quality and reliability of VLM-labeled datasets.

Building on this, we introduced a new VLM-annotated dataset, ACLImage, and conducted extensive benchmarking experiments using state-of-the-art CLL algorithms. Additionally, we conducted a dataset-level ablation study to gain deeper insights into the characteristics of VLM-annotated datasets in comparison to human-annotated datasets. Our contributions are summarized as follows:

1. We demonstrate the transformative potential of VLMs in replacing human labeling, drastically reducing costs and enabling the automated creation of large-scale datasets with efficiency.

2. To the best of our knowledge, we are the *first* to propose a unified mechanism that supports labeling scenarios in weakly-supervised learning with a robust labeling protocol to tackle challenges associated with VLM-annotated.
3. Through an extensive benchmarking study, we compare VLM-annotated with human-annotated across four datasets. We discovered the efficacy of VLM-annotated as follows:
   - Compared with existing label noise of complementary-label datasets, using VLM for labeling is better than human annotations.
   - VLM-annotated demonstrates lower bias compared to human-annotated datasets, a critical factor that significantly impacts model performance.

2 Background

2.1 Complementary-Label Learning Algorithms

In the standard supervised multi-class classification, referred to as the ordinary label learning (OLL), a labeled dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ from an unknown distribution $P(\mathbf{x}, y)$ is given, where $\mathbf{x}_i$ is the M-dimension feature vector and $y_i \in [K] = \{1, 2, \ldots, K\}$ is the class label of instance i. In CLL settings, instead of the label y_i, we have a complementary label $\bar{y}_i$ for each instance, which indicates a class that $\mathbf{x}_i$ does *not* belong to. CLL and OLL share the goal of predicting the correct labels of unseen instances.

To learn from indirect label information, researchers have made assumptions about the generation process of CLs to ensure the feasibility of learning problems and algorithms. A common assumption is the *class-conditional assumption* [4], which states that the distribution of a complementary label depends only on its ordinary label and is independent of the instance's features, i.e., $P(\bar{y}_i \mid \mathbf{x}_i, y_i) = P(\bar{y}_i \mid y_i)$ for each i. To represent the relationship between complementary and ordinary labels, a *transition matrix* is used, where $T_{j,k}$ is the probability of obtaining a complementary label k given the ordinary label j, i.e., $T_{j,k} = P(\bar{y} = k \mid y = j)$ for all $j, k \in [K]$. There are two common assumptions for CLL: (1) *Label noiseless* $T_{j,j} = 0$ requires that the generation process produce CLs from the remaining classes. If the diagonal of T is greater than zero, it is considered *label noise*, (2) *Uniform transition matrix* $T_{j,k} = \frac{1}{K-1}$ specifies that CLs are generated uniformly. If the generation process is non-uniform $T_{j,k} \neq \frac{1}{K-1}$, it is called 'bias' and represented by a *biased transition matrix.*

Building on this assumption, the previous works converted the risk minimization in OLL into an unbiased risk estimation (**URE**) [1]. The surrogate complementary loss (**SCL**) algorithm later addressed URE's overfitting issue by designing loss functions to reduce variance in the empirical estimation. To relax the two assumptions, the forward-correction loss (**FWD**) method accommodates biased transition matrix by adding a transition layer into deep neural networks to improve the estimation of the transition matrix. Alternatively, the complementary probability estimates (**CPE**) tackle biased transition matrix and label noise by reducing the CLL problem to probability estimation to remedy the impact from the transition matrix. Beyond a single complementary label per instance, the previous work studied multiple complementary label (**MCL**) problems [16]. They assumed that a label collection protocol randomly selects a label set and asks labelers whether the correct label is included. They then established the URE framework for learning with MCLs. The `libcll` benchmark [15] has showed that MCL outperforms other CLL algorithms with comprehensive ablation studies on two assumptions.

2.2 Visual-Language Models

Visual-Language Models (VLMs) have recently demonstrated remarkable capabilities in object recognition and visual reasoning. For instance, LLaVA has showcased advanced proficiency in processing images and textual prompts to generate accurate textual descriptions [17]. Building upon this foundation, LLaVA-1.6 further enhances compositional reasoning and data efficiency through a fully-connected vision-language connector [18]. Recent extensions of the LLaVA framework have expanded its applicability to video scenarios, demonstrating the versatility of VLMs across different modalities [19, 20].

With the growing maturity of VLM capabilities, their integration into data labeling processes has garnered increasing attention in computer vision (CV) tasks. For example, VLMs have been successfully adapted as labelers for downstream image recognition tasks [8] and as robust detectors to identify and mitigate noisy labels [14]. In the domain of weakly-supervised learning, VLMs have been utilized to generate strong positive and negative pseudo-labels for multi-label learning [11, 12] and to provide partial annotations [13]. The integration of VLMs into label collection protocols offers a transformative opportunity to enhance labeling efficiency and accuracy in CV tasks, marking a promising step forward in addressing challenges within data annotation workflows.

3 Complementary Label Collection Protocol for VLMs

Previous works studied CLL algorithms based on *synthetic complementary datasets*, which simplifies the problem for theoretical analysis but remains a gap in applying CLL algorithms to real-world problems. CLImage[1] [7] collected the real-world complementary label by human annotations based on CIFAR [21] and TinyImageNet200 [22] datasets.

For VLMs-based labeling, we first utilized the same complementary-labeled datasets as in CLImage [7], which were annotated by human workers on Amazon Mechanical Turk (MTurk). These datasets include complementary label data derived from CIFAR10, CIFAR20, MIN10, and MIN20. To ensure a fair comparison with the human-annotated version, we inherit the collection methodology described in [7] with VLMs. The protocol for generating CLs for each image $\mathbf{x}$ is described as follows: (1) Uniformly sample four labels without replacement from the label set $[K]$, (2) Request a VLM to select one complementary label $\bar{y}$ from the four sampled labels, (3) Add the pair $(\mathbf{x}, \bar{y})$ to the complementary dataset.

[1] `https://github.com/ntucllab/CLImage_Dataset`

`Prompt 0: Convert the CLImage's question to VLMs's prompt`

Question: <image> Please select any one "incorrect" label of this image? Answer the question by picking one wrong label: [labels[0], labels[1], labels[2], labels[3]].

Table 1: The results of prompt converting the human-annotated for VLMs on the MicroImageNet10 dataset, smaller value is better.

Platform	Model	Noisy	Time(hrs)
VLM	`LLaVA-7b-hf`	10.34%	3
MTurk	`Human`	5.19%	72

Table 2: Comparison of complementary label noise levels across different datasets with three prompts.

	CIFAR10	CIFAR20	MIN10	MIN20	Average
`Prompt 1`	**0.25%**	0.79%	0.40%	**0.29%**	**0.43%**
`Prompt 2`	0.58%	1.08%	**0.30%**	0.40%	0.59%
`Prompt 3`	0.43%	**0.78%**	0.35%	**0.29%**	0.46%

In step (2), to replicate the human annotation process, we initially adapted the question format used in CLImage [7], which asked human annotators.[2] This question was converted into a `Prompt 0` for VLMs to annotate CLs. We discovered that VLM-annotated datasets demonstrate inferior effectiveness compared to human-annotated datasets, exhibiting higher label noise while requiring significantly less time to complete the labeling task, as shown in Table 1[3]. Additionally, we hypothesize the label noise of VLMs-annotated could be reduced via optimization prompting since VLMs are impacted by corresponding lexical changes with the same semantic sentences [23, 24]. Therefore, we developed tailored prompts to reduce label noise rates for VLMs-based labeling. Through extensive testing, we identified three optimal ones (`Prompt 1, 2, 3`) for complementary labeling tasks. These prompts were carefully adapted from the instructions used during the model's training [20]. Among them, `Prompt 1` achieved the best performance, minimizing label noise rate cross all datasets, as shown in Table 2. Consequently, `Prompt 1` was selected to handle the labeling tasks. The detailed of labeling framework using VLMs for complementary-label is illustrated in Figure 1.

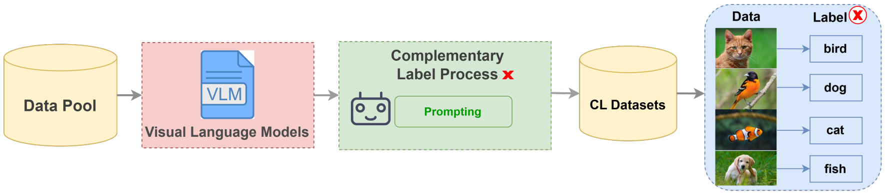


Fig. 1: Label Collection Protocol Using Visual Language Models for Complementary Labels (CLs).

[2] Please select any one *incorrect* label for this image? Pick one wrong label: [choice[1], choice[2], choice[3], choice[4]].

[3] We reached out to the authors of CLImage regarding the time required for human labeling, they confirmed that it took approximately 3 days to annotate each dataset.

Prompt 1

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

Prompt 2

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.

Prompt 3

Question: <image> Which label does not belong to this image? (1) labels[0] (2) labels[1] (3) labels[2] (4) labels[3] Answer with the given number directly.

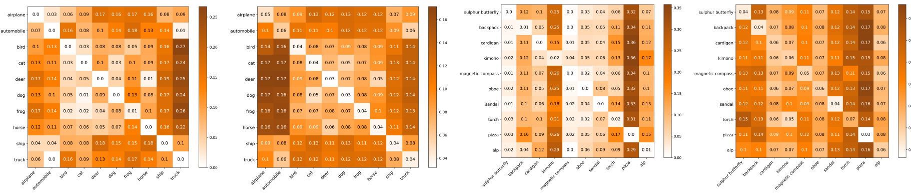
Markedly, the label noise rate on the MIN10 dataset was reduced by more than *25 times*, dropping from 10.34% to 0.40%. These findings highlight the critical importance of prompt design in improving labeling accuracy and reliability in VLMs-based annotation processes.

4 Dataset Characteristic

In this section, we analyze the CLs collected through VLM-based labeling. Specifically, we compare the label noise rate, label distribution, and transition matrix between human-labeled and VLM-labeled datasets. These comparisons provide deeper insights into the differing behaviors between the human annotation process and the VLM labeling protocol.

Characteristic 1: low label noise rate Our collected VLM-labeled CLs exhibit significantly lower label noise rates compared to human-labeled datasets. For human-labeled datasets, the average label noise rates are 3.39% for CLCIFAR10, 2.80% for CLCIFAR20, 5.19% for CLMIN10, and 3.21% for CLMIN20 [7]. In contrast, the noise rates for VLM-labeled datasets are remarkably lower: 0.24% for ACLCIFAR10, 0.89% for ACLCIFAR20, 0.66% for ACLMIN10, and 0.86% for ACLMIN20. These label noise rates are the average of three CLs. Figure 5a provides a visual comparison of label noise rates between VLM- and human-labeled datasets. These results underscore the superior performance of VLM annotators in complementary-label annotation tasks, achieving significantly lower noise rates.

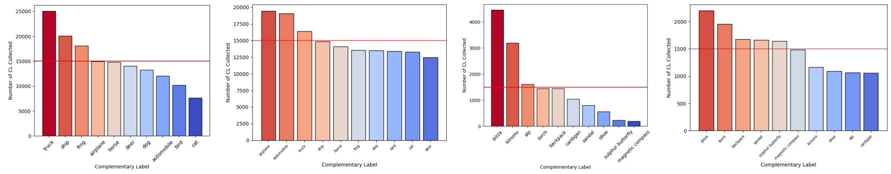
Characteristic 2: highly biased transition matrix Figure 2 presents the empirical transition matrices of CLs for CIFAR10 and MIN10 datasets under both VLM-labeled and human-labeled protocols. Upon examining the figure, we observe that the transition matrix of the VLM-labeled datasets exhibits a higher degree of bias compared to the human-labeled datasets. This bias aligns with our findings from "Characteristic 3", where we identified a significantly higher imbalance ratio in the VLM-labeled datasets. This trend extends to the CIFAR20 and MIN20 datasets as well, although for brevity, only CIFAR10 and MIN10 are shown here.



(a) ACLCIFAR10 (b) CLCIFAR10 (c) ACLMIN10 (d) CLMIN10

Fig. 2: The empirical transition matrices of complementary datasets ACLCIFAR10 vs. CLCIFAR10, and ACLCIFAR10 vs. CLCIFAR10.

Characteristic 3: highly imbalanced label distribution Figure 3 illustrates a different level of imbalance distribution of CLs in VLM-labeled datasets compared to their human-labeled counterparts. Specifically, we observe that the VLM-labeled datasets are significantly more imbalanced. For example, the imbalance ratio for human-labeled datasets is approximately 1.56 and 2.07 for CLCIFAR10 and CLMIN10, respectively. In contrast, the imbalance ratio for VLM-labeled datasets increases dramatically to around 3.26 for ACLCIFAR10 and 23.46 for ACLMIN10. Further analysis reveals differing biases in the preferred categories between the two labeling protocols. For the CIFAR10 dataset, human-labeled data tends to favor categories such as “airplan” and “automobile”, whereas VLM-labeled data shows a preference for “truck” and “ship”. Similarly, in the MIN10 dataset, both human and VLM labeling protocols show a preference for categories like “pizza” and “kimono”. However, their contrasting biases are evident, as human-labeled data leans toward “sulphur butterfly” and “magnetic compass”, while VLM-labeled data favors “alp” and “cardigan”. These findings highlight distinct behaviors between human and VLM-based complementary labeling protocols, shedding light on their unique biases and distribution patterns.



(a) ACLCIFAR10 (b) CLCIFAR10 (c) ACLMIN10 (d) CLMIN10

Fig. 3: The label distribution of ACLCIAFR10, CLCIFAR10, ACLMIN10, and CLMIN10 datasets.

Notably, these patterns are consistent across all four VLM-labeled datasets when compared to their human-labeled counterparts. While VLM-labeled datasets demonstrate a lower label noise rate, they exhibit greater imbalances in label distribution and higher levels of bias. These findings emphasize the trade-offs inherent in using VLM for complementary labeling tasks. In the next section, we validate our methodology to assess the practicality and reliability of collecting CLs in real-world scenarios.

5 Experiments

In our experiments, we evaluated the effectiveness of VLMs as replacements for human annotators in labeling CLs, namely, ACLImage[4]. We observed notable accuracy performance gaps between human-labeled and VLM-labeled datasets, and VLMs' potential to reduce labeling costs in CLL. Further analyses revealed that biased transition matrices are major challenges, which we addressed through label cleaning and a newly designed protocol, achieving promising results. Additionally, scalability experiments demonstrated that VLMs perform effectively in large-scale settings, marking significant progress toward practical, cost-efficient CLL.

5.1 Experimental Setups

The experiments presented in Sections 5.2 and 5.4 evaluate six CLL algorithms selected for their strong performance on CLImage datasets with three CLs, as reported in the `libcll` benchmark. The selected algorithms include SCL-NL [6], SCL-EXP [6], MCL-LOG [16], FWD [4], CPE-F [25], and CPE-T [25]. These algorithms were tested on VLM-annotated datasets and compared with corresponded human annotation datasets, using three CLs per instance in all cases.

To ensure consistency, we adopted hyperparameters established in `libcll`. Specifically, training was performed with a fixed batch size of 256 for 300 epochs on NVIDIA Tesla V100 GPUs with 32GB memory. Learning rates were selected based on the best accuracies from the set `{1e-3, 5e-4, 1e-4, 5e-5, 1e-5}`. All models were trained using the Adam optimizer and a ResNet34 backbone. For each experiment, 10% of the training data was reserved as a validation set, assuming access to ordinary labels to calculate validation accuracy and conducted four trials with different random seeds to ensure robustness.

Table 3: Performance comparison of different CLL algorithms on Human-labeled and VLM-labeled datasets.

	ACLCIFAR10	CLCIFAR10	ACLCIFAR20	CLCIFAR20	ACLMIN10	CLMIN10	ACLMIN20	CLMIN20
SCL-NL	**53.37**±0.50	**47.30**±0.50	5.23±0.39	8.59±0.75	16.21±0.62	12.87±2.33	9.39±0.87	6.87±0.39
SCL-EXP	33.20±3.68	47.12±0.91	5.65±0.79	**9.74**±0.52	16.16±1.21	12.78±1.42	8.82±0.48	**7.10**±0.83
MCL-LOG	52.79±0.25	46.13±0.57	**6.35**±0.22	8.57±0.20	**16.57**±0.75	**15.02**±2.25	**12.57**±1.08	6.55±0.95
FWD	**69.49**±1.16	**52.48**±0.63	**33.39**±0.29	**24.56**±0.95	**49.42**±3.56	29.33±0.85	**28.51**±1.13	**10.11**±1.29
CPE-F	69.10±1.11	51.74±0.98	33.25±0.30	24.44±1.07	48.98±3.05	**29.51**±0.95	27.19±1.34	9.52±1.71
CPE-T	62.43±1.21	49.79±1.45	19.08±0.69	20.85±0.52	43.00±2.37	27.97±1.06	22.48±1.89	9.70±1.13
Supervision	86.61±0.30		64.46±0.72		66.64±1.00		60.04±1.97	

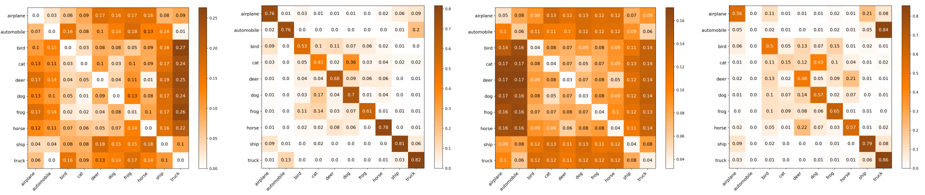
To facilitate a fair comparison, we categorized algorithms based on their use of a transition matrix, which some algorithms require to compute the loss. Algorithms leveraging a transition matrix, referred to as T-aware, are listed in the upper sections of the result tables, while T-agnostic algorithms are presented in the lower sections. This separation highlights the differences in their performance and computational frameworks.

[4] `https://github.com/yahcreepers/PAKDD_ACLImage_Dataset`

5.2 Standard Benchmark on ACLImage

In this section, we present the baseline learning results on our VLM-annotated datasets and directly compare them with those obtained from human-annotated datasets. The results are summarized in Table 3. Notably, VLM-annotated datasets improved the performance of most algorithms. However, some T-agnostic methods exhibited decreased performance on VLM-annotated datasets. We hypothesize that this decline is due to the inherent bias in the VLM-annotated datasets. T-agnostic methods operate under the assumption of a uniform distribution and are unable to adapt to shifts in the complementary-label distribution. Consequently, they are more susceptible to deviations in the transition matrices caused by dataset biases.

The relationship between transition matrices and testing accuracy is illustrated in Figure 4. The figure underscores the significant impact of variations in label noise rate and bias on model performance. Specifically, higher label noise rates and larger deviations in the transition matrix introduce ambiguity, leading to substantial performance degradation. This finding, combined with Figure 5a reveals a critical limitation of human-annotated, which is prone to high label noise rates that can severely impair the effectiveness of current CLL algorithms. In contrast, VLM-annotated demonstrates a clear advantage in achieving significantly lower label noise rates. These results emphasizes the need to develop robust methods to mitigate the impact of biased transition matrices.



(a) ACLCIFAR10 Transition Matrix

(b) ACLCIFAR10 Testing Accuracy

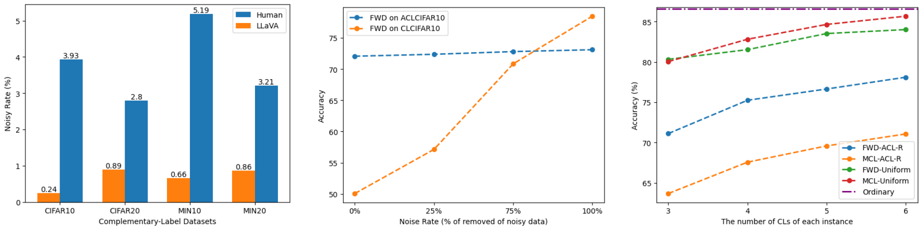
(c) CLCIFAR10 Transition Matrix

(d) CLCIFAR10 Testing Accuracy

Fig. 4: The comparison between FWD predictions learned from human-labeled and VLM-labeled datasets.

5.3 Label Noise Removal

In this section, our aim was to isolate the impact of noisy labels and identify the main cause of the accuracy performance gap between human and VLM-annotated datasets. To achieve this, we measured the accuracy on noise-reduced versions of the CLCIFAR10 and ACLCIFAR10 datasets, progressively removing different proportions (0%, 25%, 75%, and 100%) of noisy labels. The results, presented in Figure 5b, indicate that while VLM-annotated datasets exhibit a lower label noise rate, human-annotated CLImage datasets outperform VLM-annotated datasets when noisy labels are sufficiently reduced. This observation underscores the critical role of the label noise rate in the quality of CLs and



(a) Complementary label noise levels across datasets. (b) Result of noisy label cleaning on ACLCIFAR10. (c) Result of incremental CLs on ACLCIFAR10.

Fig. 5: The ablation study of comparison between human-labeled datasets vs VLM-labeled datasets.

reinforces the value of VLM labeling as a means to facilitate CLL in real-world scenarios. These findings suggest the need to reduce the bias in collected labels and improve learning methods to effectively address distributional biases.

5.4 Bias Removal

To reduce bias in collected labels, we developed a new protocol tailored for VLM-based annotation. The protocol involves the following steps: (1) creating a weight list of label distributions and candidate sets for each instance: $W_{\bar{y}_i}$, and (2) iteratively collecting CLs through the following process: (i) weighted sampling of four labels from the candidate set K_i, (ii) prompting the VLM to select one complementary label from the sampled labels $\bar{y}_i$ from the sampled labels, (iii) adding the pair $(\mathbf{x_i}, \bar{y}_i)$ to the complementary dataset, (iv) removing $\bar{y}_i$ from the candidate set K_i, and (v) reducing the weight $W_{\bar{y}_i}$ by one—the process will stop when weight $W_{\bar{y}_i}$ is equal the sampling number of candidate set. Using this protocol, we labeled CIFAR10 and MIN10, resulting in the ACLCIFAR10-R and ACLMIN10-R datasets. This protocol successfully mitigates biases inherent in VLM-annotated labels and improves learning outcomes on the newly collected datasets. As illustrated in Table 4 and Figure 6, this carefully designed labeling process significantly reduces bias in VLM annotations while lowering label noise rates. These results underscore the potential of VLM-annotated for complementary labeling, enhancing their applicability in weakly supervised learning scenarios and demonstrating the efficacy of structured, bias-aware label collection protocols.

Table 4: Performance of CLL algorithms with reducing biasedness approach.

	ACLCIFAR10-**R**	ACLCIFAR10	ACLMIN10-**R**	ACLMIN10
SCL-NL	<u>64.46</u>±0.48	**53.37**±0.50	<u>24.72</u>±0.70	<u>16.21</u>±0.62
SCL-EXP	62.35±0.73	33.20±3.68	21.98±0.77	16.16±1.21
MCL-LOG	**64.92**±0.52	<u>52.79</u>±0.25	**27.38**±3.03	**16.57**±0.75
FWD	**71.03**±0.44	**69.49**±1.16	<u>51.29</u>±2.64	**49.42**±3.56
CPE-F	<u>70.81</u>±0.08	<u>69.10</u>±1.11	**51.30**±2.91	<u>48.98</u>±3.05
CPE-T	63.56±0.53	62.43±1.21	47.76±1.26	43.00±2.37

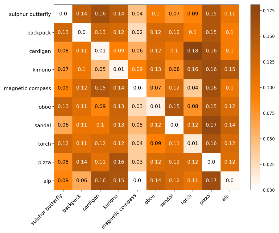


Fig. 6: ACLMIN10-R

5.5 Incremental Complementary Labels (CLs)

In this section, we investigate the potential of increasing the number of CLs per instance to reduce the performance gap between CLL and OLL. Building upon the labeling protocol introduced in Section 5.4, we extended the number of CLs to six per instance. This experiment highlights a key advantage of VLMs: their capacity to generate a substantial volume of labels at a significantly lower cost, thereby enhancing the practicality of CLL. Additionally, we compared the outcomes of bias removal with those of uniform synthetic datasets to evaluate the negative effects of bias in real-world complementary datasets and quantify the performance gaps. As depicted in Figure 5c, increasing the number of CLs per instance results in improved model performance. However, notable gaps persist between CLL and OLL, as well as between VLM-annotated datasets and ideally uniform datasets. These findings emphasize the critical challenge of addressing biases in collected CLs. They further suggest that simply increasing the number of CLs is insufficient to bridge these gaps, underscoring the need for innovative strategies to mitigate labeling biases effectively.

6 Conclusion

In this paper, we introduced a novel VLM-based protocol for collecting CLs, addressing label noise and bias that arise when adapting human annotation processes. Our optimized framework significantly improved labeling quality, enabling the creation of ACLImage, the first publicly available, VLM-annotated complementary-label dataset. Our experiments demonstrated that VLM-annotated datasets effectively reduce label noise and bias while achieving competitive performance across various CLL algorithms. By minimizing dependence on costly and error-prone human annotations, our work advances the practicality and scalability of CLL and establishes a foundation for leveraging VLMs in weakly-supervised learning.

References

1. Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. Learning from complementary labels. In *Proceedings of the 31st NeurIPS*, page 5644–5654, 2017.
2. Takashi Ishida, Gang Niu, and Masashi Sugiyama. Binary classification from positive-confidence data. In *NeurIPS*, 2018.
3. Takashi Ishida, Gang Niu, Aditya Menon, and Masashi Sugiyama. Complementary-label learning for arbitrary losses and models. In *Proceedings of the 36th ICML*, pages 2971–2980, 2019.
4. Xiyu Yu, Tongliang Liu, Mingming Gong, and Dacheng Tao. Learning with biased complementary labels. In *Computer Vision – ECCV 2018*, pages 69–85, 2018.
5. Yuzhou Cao, Shuqi Liu, and Yitian Xu. Multi-complementary and unlabeled learning for arbitrary losses and models. *Pattern Recognition*, 124:108447, 2022.
6. 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 *Proceedings of the 37th ICML*, pages 1929–1938, 2020.

7. Hsiu-Hsuan Wang, Tan-Ha Mai, Nai-Xuan Ye, Wei-I Lin, and Hsuan-Tien Lin. Climage: Human-annotated datasets for complementary-label learning, 2023.
8. Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.
9. Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. 2023.
10. Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 2023.
11. Xin Xing, Zhexiao Xiong, Abby Stylianou, Srikumar Sastry, Liyu Gong, and Nathan Jacobs. Vision-language pseudo-labels for single-positive multi-label learning. In *Proceedings of the IEEE/CVPR*, pages 7799–7808, 2024.
12. Anonymous. Weak supervision from vision-language models to self-improve on downstream tasks. In *Submitted to The Thirteenth International Conference on Learning Representations*, 2024. under review.
13. Qian-Wei Wang, Yuqiu Xie, Letian Zhang, Zimo Liu, and Shu-Tao Xia. Pre-trained vision-language models as partial annotators. 2024.
14. Tong Wei, Hao-Tian Li, Chun-Shu Li, Jiang-Xin Shi, Yu-Feng Li, and Min-Ling Zhang. Vision-language models are strong noisy label detectors. In *Advances in Neural Information Processing Systems 37*, 2024.
15. 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.
16. Lei Feng, Takuo Kaneko, Bo Han, Gang Niu, Bo An, and Masashi Sugiyama. Learning with multiple complementary labels. In *Proceedings of the 37th ICML*, ICML'20, 2020.
17. Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
18. Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVPR*, 2024.
19. Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.
20. Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024.
21. Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.
22. Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.
23. Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.
24. Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. Sensitivity of generative vlms to semantically and lexically altered prompts. *arXiv preprint arXiv:2410.13030*, 2024.
25. Wei-I Lin and Hsuan-Tien Lin. Reduction from complementary-label learning to probability estimates. In *Proceedings of the PAKDD*, May 2023. winner of the best paper runner-up award.