# Mitigating Endogenous Confirmation Bias in Noisy Label Learning for Vision-Language Models

**Feiyang Ning, Xinyang Chen***

School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), China
{ningfeiyang2002, chenxinyang95}@gmail.com

## Abstract

Pretrained vision-language models (VLMs), especially CLIP, excel at adapting to downstream tasks through fine-tuning with sufficient high-quality labeled data. However, real-world training data often contains noisy labels, leading to significant performance degradation when models are naively fine-tuned on them. Existing noisy label learning methods for VLMs typically leverage the model's own pretrained knowledge, either via zero-shot predictions or vanilla self-training based on them, to identify and handle noisy samples. Crucially, these approaches blindly trust the VLM's pretrained knowledge, which can introduce endogenous confirmation bias: erroneous pretrained priors lead to incorrect noise detection, further amplifying the bias and corrupting the model. To overcome this limitation, we propose the **D**ebiased **K**nowledge **A**daptation **F**ramework (**DKAF**), which empowers the model to challenge and correct potentially flawed zero-shot predictions. DKAF operates in three progressive phases: (1) **Clean Sample Selection**. We introduce a cross-modal collaborative pseudo-labeling to train a robust noisy label detector, explicitly mitigating confirmation bias by aggregating diverse signals beyond the model's initial zero-shot view. (2) **Noisy Label Refinement**. For samples identified as noisy, we apply a dual-modal consistency strategy to selectively correct their labels, leveraging alignment between dominant and fused modalities to guide refinement while minimizing reliance on potentially biased internal knowledge. (3) **Model Adaptation**. The model is progressively fine-tuned using the jointly curated dataset of selected clean samples and corrected noisy samples, promoting robust adaptation to the target task. Extensive experiments on nine benchmark datasets (both synthetic and real-world noise) demonstrate that DKAF consistently outperforms state-of-the-art multimodal noisy label learning methods. Notably, under high-noise conditions, DKAF achieves average accuracy improvements of 3.08%.

**Code** — https://github.com/Feiyang-Ning/DKAF

## 1 Introduction

Vision-language models (VLMs), such as CLIP (Radford et al. 2021), demonstrate strong generalization capabilities through large-scale multimodal pre-training. To optimize performance on downstream tasks, fine-tuning with task-specific data remains essential (Zhou et al. 2022b; Gao et al. 2024). However, the effectiveness of fine-tuning heavily relies on high-quality labeled data for the target task. In practice, the training data often contains label noise due to human annotation errors and weakly supervised labeling mechanisms (Zhou 2018). Directly fine-tuning pre-trained models on noisy datasets may lead to performance degradation.

Predominant learning with noisy labels (LNL) approaches comprise three categories: sample selection, robust loss, and label correction. Among these, sample selection is the most widely adopted approach. It focuses on identifying clean samples based on specific criteria and fine-tuning pre-trained models with them. Most of such methods rely on the small-loss strategy, which assumes that samples with smaller training losses are more likely to be clean (Jiang et al. 2018; Han et al. 2018; Shen and Sanghavi 2019). However, this strategy is inherently sensitive to the type and distribution of la-

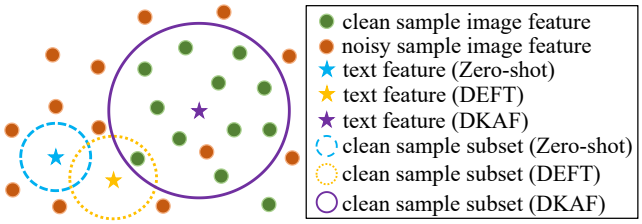



Figure 1: Recognition issues on hard classes. Hard classes tend to cause misalignment in zero-shot features, whereas DKAF effectively selects higher-quality clean samples.

| Method | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|
| Bombay (Dataset: OxfordPets, Noise Rate: 40%) | | | |
| Zero-shot | 0.00 | 0.00 | 0.00 |
| DEFT | **100.00** | 8.26 | 15.87 |
| **DKAF (Ours)** | 98.31 | **100.00** | **99.15** |
| Whip-poor-will (Dataset: CUB-200-2011, Noise Rate: 40%) | | | |
| Zero-shot | 0.00 | 0.00 | 0.00 |
| DEFT | **100.00** | 5.26 | 9.99 |
| **DKAF (Ours)** | **100.00** | **100.00** | **100.00** |

Table 1: Recognition results on representative hard classes.

*Corresponding author.

bel noise (Song et al. 2022). In particular, it may mistakenly select noisy samples with small losses and exclude hard but clean samples with large losses, thereby reducing the effectiveness of clean sample selection.

Benefiting from the advancements of VLMs, especially CLIP, a few noisy label learning methods based on VLMs have been proposed to perform clean sample selection by leveraging their strong cross-modal semantic understanding capacity. Specifically, CLIPCleaner (Feng, Tzimiropoulos, and Patras 2024) and GSM (Liang et al. 2025) construct a CLIP-based zero-shot classifier for clean sample selection. DEFT (Wei et al. 2024) adopts vanilla self-training by learning dual textual prompts to detect noisy labels.

Compared to visual approaches, these multimodal methods better exploit image-text semantic correlations to enhance sample selection. However, they suffer from endogenous confirmation bias due to over-reliance on CLIP's pretrained knowledge. As shown in Tab.1, even state-of-the-art methods, DEFT (Wei et al. 2024), fail to rectify samples with erroneous CLIP zero-shot predictions. This bias manifests as preferential treatment toward overestimated classes (Huang, Chu, and Wei 2022; Alabdulmohsin et al. 2024), causing repeated selection of such samples while neglecting hard classes, ultimately degrading clean sample quality.

To advance noisy label learning with CLIP, we propose the Debiased Knowledge Adaptation Framework (DKAF), which empowers the model to challenge and correct potentially flawed zero-shot predictions. Fig.1 illustrates why zero-shot and DEFT perform poorly in recognizing hard classes: CLIP lacks sufficient learning of these classes during the pretraining phase, which limits its semantic alignment capability and prevents it from accurately representing the visual concepts of the target classes, ultimately resulting in ineffective predictions. In contrast, DKAF mitigates overreliance on pretrained knowledge through three phases: (1) Clean Sample Selection: Cross-modal collaborative pseudo-labeling trains textual prompts by aggregating visual-textual predictions for robust noise detection; (2) Noisy Label Refinement: A dual-modal consistency strategy selectively corrects labels using inter-modal alignment, reducing dependence on biased priors; (3) Model Adaptation: Downstream adaptation leverages curated clean and refined samples. By addressing endogenous confirmation bias, DKAF achieves precise visual-concept alignment. Tab.1 demonstrates significant performance gains on hard-class recognition.

We conducted extensive experiments on six synthetic and three real-world datasets, demonstrating that DKAF outperforms existing multimodal methods in noisy label learning. Our main contributions can be summarized as follows:

- We critically reveal that excessive reliance on CLIP's zero-shot capability in existing methods propagates uncorrected errors, inducing endogenous confirmation bias.
- To mitigate confirmation bias, we propose cross-modal collaborative pseudo-labeling to train textual prompts for noise detection. This method aggregates predictions from both textual and visual views to produce reliable pseudo-labels. A dual-modal consistency strategy is then employed to selectively refine noisy labels, reducing dependence on potentially biased inherent priors.
- We conduct experiments on nine benchmark datasets and achieve an average accuracy improvement of 3.08% under high-noise conditions over the SOTA methods.

## 2 Related Work

**Prompt Tuning** In recent years, vision-language models (VLMs) such as CLIP, ALIGN (Jia et al. 2021), and BLIP (Li et al. 2022) have demonstrated strong multimodal representation capabilities. Prompt tuning leverages learnable prompts to adapt VLMs to downstream tasks, instead of modifying backbone weights (Bordes et al. 2024). Most existing methods focus exclusively on the textual branch by introducing learnable prompt vectors, such as CoOp (Zhou et al. 2022b), CoCoOp (Zhou et al. 2022a) and PLOT (Chen et al. 2022). Some methods, like VPT (Jia et al. 2022), instead insert a few trainable parameters into the visual encoder's input space. Others, such as MaPLe (Khattak et al. 2023), apply learnable prompts to both the textual and visual branches. Our method draws on the idea of prompt learning to select clean samples, and leverages prompt tuning to enhance CLIP's performance on downstream tasks.

**Learning with Noisy Labels** Abundant methods have been proposed to address noisy label learning, which can be categorized into three types: (1) Sample selection methods focus on identifying clean samples from noisy-labeled datasets. Representative methods include Co-teaching (Han et al. 2018), JoCoR (Wei et al. 2020) and GMM (Li, Socher, and Hoi 2020). (2) Robust loss methods aim to design noise-tolerant loss functions. Representative methods include GCE (Zhang and Sabuncu 2018), SCE (Wang et al. 2019) and ELR (Liu et al. 2020). (3) Label correction methods aim to assign accurate pseudo-labels to noisy samples. Representative methods include Bootstrapping (Reed et al. 2014) and PENCIL (Yi and Wu 2019). Additionally, several methods have recently applied VLMs such as CLIP for learning with noisy labels: CLIPCleaner (Feng, Tzimiropoulos, and Patras 2024), GSM (Liang et al. 2025) and DEFT (Wei et al. 2024). Unlike prior multimodal methods that use only zero-shot prediction or vanilla self-training, our method focuses on enhancing CLIP's sample selection capability by mitigating its endogenous confirmation bias, and selectively corrects noisy labels through a dual-modal consistency strategy, thereby improving performance in noisy label learning.

## 3 Methodology

This paper addresses the challenge of robustly adapting CLIP for learning with noisy labels. Formally, consider a noisy-labeled training dataset $\mathcal{D} = (\mathcal{X}, \mathcal{Y}) = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathcal{X}$ denotes an image and $y_i \in \mathcal{Y}$ is its potentially misannotated label over $C$ classes ($\mathcal{Y} = \{1, \ldots, C\}$). Critically, the underlying true labels remain unobserved.

Existing CLIP-based noisy label learning methods typically exploit the model's zero-shot capability to identify label errors: CLIPCleaner (Feng, Tzimiropoulos, and Patras 2024) and GSM (Liang et al. 2025) directly utilize zero-shot predictions for noise detection, while DEFT (Wei

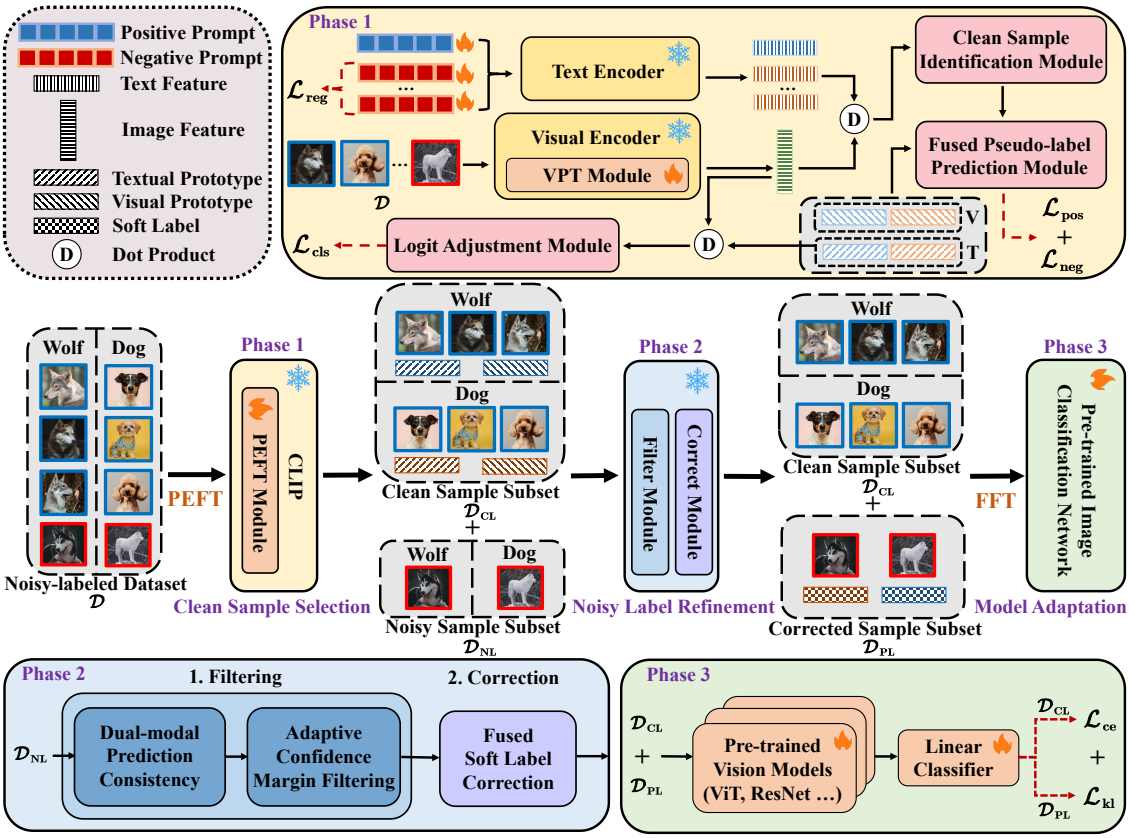



Figure 2: The architecture of the proposed framework for noisy label learning with CLIP, consisting of three progressive phases.

et al. 2024) advances this via vanilla self-training with dual prompts initialized from zero-shot outputs. However, these approaches exhibit a key limitation: they uncritically trust CLIP's pretrained knowledge, inducing confirmation bias. When this knowledge is flawed, errors in sample selection self-reinforce, amplifying bias and misleading the model.

### 3.1 Overview

To mitigate confirmation bias, we propose the Debiased Knowledge Adaptation Framework (DKAF), designed to empower CLIP to critically evaluate and rectify potentially erroneous zero-shot predictions for robust adaptation to noisy label downstream tasks. As illustrated in Fig.2, DKAF operates in three progressive phases. The first phase is **Clean Sample Selection**: We identify clean samples via self-training with multi-prompt learning. Pseudo-labels are generated by aggregating visual and textual signals beyond the initial zero-shot view, followed by logit adjustment loss to address imbalanced pretrained prediction distributions. The second phase is **Noisy Label Refinement**: To leverage information from noisy samples, we select correctable samples through dual-modal prediction consistency and adaptive confidence margin filtering, then assign fused soft labels to these samples. The third phase is **Model Adaptation**: The pre-trained deep neural network is progressively fine-tuned using both clean and corrected samples to enhance performance on downstream image classification tasks.

### 3.2 Phase 1: Clean Sample Selection

Accurate clean sample selection is pivotal for noisy label learning. CLIP inherently exhibits strong zero-shot capability, allowing direct label inference for unseen images without additional training, which facilitates efficient noisy sample detection. Advanced methods like DEFT (Wei et al. 2024) further enhance CLIP's selection ability through self-training with dual textual prompts. Specifically, for each class $k \in \mathcal{Y}$, a pair of text prompts consisting of a positive prompt $\mathrm{t}_k^+$ and a negative prompt $\mathrm{t}_k^-$ is constructed. The positive prompt serves as an anchor for the representation of the target class $k$, while the negative prompt acts as a learnable semantic threshold to distinguish clean and noisy samples. For each sample $(\mathbf{x}_i, y_i) \in \mathcal{D}$, use CLIP's visual encoder $\phi_\theta$ to extract its image feature $\boldsymbol{I}_i$. If the similarity between $\boldsymbol{I}_i$ and the positive prompt feature $\boldsymbol{T}_{y_i}^+$ exceeds that with the negative prompt feature $\boldsymbol{T}_{y_i}^-$, the sample is considered clean.

However, its pseudo-labeling strategy remains identical to vanilla self-training: it relies solely on image-text similarity scores. Crucially, since dual-prompt training is initialized from zero-shot similarity scores without additional guidance, this approach inherits and amplifies confirmation bias. To address the above limitation in sample selection, we propose a novel cross-modal collaborative pseudo-labeling strategy to mitigate confirmation bias. Subsequently, we first present how to use textual prompt learning for clean

sample identification, followed by cross-modal collaborative pseudo-labeling designed to address confirmation bias.

**Textual Prompt Learning** The learning of dual textual prompts relies on a single negative prompt to identify clean samples, making the discrimination results easily affected by individual outlier samples and incidental prompt bias. Inspired by the idea of multi-view learning (Xu, Tao, and Xu 2013), we construct a set of negative prompts $\{\mathrm{t}_k^{-(j)}\}_{j=1}^{N_{\text{neg}}}$ for each class $k \in \mathcal{Y}$ to capture multifaceted distributions of noisy-labeled samples. Formally, given a sample $(\mathbf{x}_i, y_i) \in \mathcal{D}$, the clean sample subset $\mathcal{D}_{\text{CL}}$ is defined as:

$$\mathcal{D}_{\text{CL}} = \left\{(\mathbf{x}_i, y_i) \mid \text{sim}(\boldsymbol{I}_i, \boldsymbol{T}_{y_i}^{+}) > \text{sim}(\boldsymbol{I}_i, \boldsymbol{T}_{y_i}^{-})\right\}, \quad (1)$$

where $\text{sim}(\boldsymbol{I}_i, \boldsymbol{T}_{y_i}^{-}) = \frac{1}{N_{\text{neg}}} \sum_{j=1}^{N_{\text{neg}}} \text{sim}(\boldsymbol{I}_i, \boldsymbol{T}_{y_i}^{-(j)})$.

To further enable end-to-end differentiable optimization, the binary hard discrimination criterion for identifying clean samples can be transformed into a soft formulation through continuous probability modeling:

$$p_{i,k}^{\text{clean}} = \frac{e^{\text{sim}(\boldsymbol{I}_i, \boldsymbol{T}_k^{+})/\tau}}{e^{\text{sim}(\boldsymbol{I}_i, \boldsymbol{T}_k^{+})/\tau} + e^{\text{sim}(\boldsymbol{I}_i, \boldsymbol{T}_k^{-})/\tau}}, \quad (2)$$

where $p_{i,k}^{\text{clean}}$ represents the probability of $i$-th sample belonging to the class $k$ as a clean sample. Obviously, when $p_{i,y_i}^{\text{clean}} > \frac{1}{2}$, the $i$-th sample can be considered clean.

Since the negative prompts $\{\mathrm{t}_k^{-(j)}\}_{j=1}^{N_{\text{neg}}}$ for each class $k$ are randomly initialized, we employ a regularization term $\mathcal{L}_{\text{reg}}$ to ensure their diversity. This promotes comprehensive modeling of multifaceted distributions in noisy-labeled samples. The regularization loss is defined as:

$$\mathcal{L}_{\text{reg}} = \frac{1}{C} \cdot \frac{1}{\binom{N_{\text{neg}}}{2}} \sum_{k=1}^{C} \sum_{1 \le i < j \le N_{\text{neg}}} \text{sim}(\boldsymbol{T}_k^{-(i)}, \boldsymbol{T}_k^{-(j)}), \quad (3)$$

where $\binom{N_{\text{neg}}}{2}$ denotes the number of unique prompt pairs.

**Cross-modal Collaborative Pseudo-labeling** Following Eq.(2), clean and noisy samples are constructed for optimization with respect to the positive and negative prompts using vanilla self-training. Specifically, the positive prompt loss $\mathcal{L}_{\text{pos}}$ and negative prompt loss $\mathcal{L}_{\text{neg}}$ are defined as:

$$\mathcal{L}_{\text{pos}} = \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}} -\log(p_{i,\hat{y}_i}^{\text{clean}}), \quad (4)$$

$$\mathcal{L}_{\text{neg}} = \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}} -\log(1 - p_{i,\bar{y}_i}^{\text{clean}}), \quad (5)$$

where $\hat{y}_i$ is the pseudo-label of the $i$-th sample (if the sample is considered clean in this epoch, then $\hat{y}_i = y_i$; otherwise, $\hat{y}_i$ is the label predicted by the current CLIP), and $\bar{y}_i$ is the negative label (randomly selected from $\mathcal{Y}$ such that $\bar{y}_i \neq y_i$).

The core limitation of the above vanilla pseudo-labeling strategies stems from their reliance on a single generation method: pseudo-labels are determined exclusively by image-text similarity in the feature space. This approach is problematic due to the inherent semantic gap between visual and textual modalities in CLIP, which compromises pseudo-label accuracy (Xing et al. 2023). When a domain gap exists between CLIP's pretraining dataset and the target dataset, this semantic misalignment intensifies, degrading zero-shot prediction quality in downstream tasks.

To overcome this constraint, we incorporate visual prototype to calibrate pseudo-label generation. Concretely, after each training epoch, construct the visual prototype $\boldsymbol{V}_k$ for each class $k$ using the detected $\mathcal{D}_{\text{CL}}$ from this epoch:

$$\boldsymbol{V}_k = \frac{1}{|\mathcal{D}_k|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_k} \boldsymbol{I}_i, \quad (6)$$

where $\mathcal{D}_k = \{(\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{CL}} \mid y_i = k\}$.

Then, for each noisy sample $(\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{NL}}$, we compute its similarity-based probability distributions with respect to all positive textual prompt features and visual prototypes:

$$\boldsymbol{P}_i^{\text{L}} = \{p_i^{\text{L}}(c)\}_{c=1}^{C}, \quad p_i^{\text{L}}(c) = \frac{e^{\text{sim}(\boldsymbol{I}_i, \boldsymbol{T}_c^{+})/\tau}}{\sum_{k=1}^{C} e^{\text{sim}(\boldsymbol{I}_i, \boldsymbol{T}_k^{+})/\tau}}; \quad (7)$$

$$\boldsymbol{P}_i^{\text{V}} = \{p_i^{\text{V}}(c)\}_{c=1}^{C}, \quad p_i^{\text{V}}(c) = \frac{e^{\text{sim}(\boldsymbol{I}_i, \boldsymbol{V}_c)/\tau}}{\sum_{k=1}^{C} e^{\text{sim}(\boldsymbol{I}_i, \boldsymbol{V}_k)/\tau}}. \quad (8)$$

Finally, by integrating $\boldsymbol{P}_i^{\text{L}}$ and $\boldsymbol{P}_i^{\text{V}}$, the fused probability distribution $\boldsymbol{P}_i^{\text{M}}$ of the pseudo-label is obtained:

$$\boldsymbol{P}_i^{\text{M}} = \{p_i^{\text{M}}(c)\}_{c=1}^{C}, \ p_i^{\text{M}}(c) = \alpha \cdot p_i^{\text{L}}(c) + (1-\alpha) \cdot p_i^{\text{V}}(c), \quad (9)$$

where the pseudo-label is assigned as $\hat{y}_i = \arg\max_c(p_i^{\text{M}}(c))$. $\alpha$ denotes an adaptive weight quantifying the reliability of both textual class centers and visual prototypes. Specially, $\alpha$ is updated after each epoch based on samples in $\mathcal{D}_{\text{CL}}$:

$$\alpha = \frac{N_{\text{text}}}{N_{\text{text}} + N_{\text{vision}}}, \quad (10)$$

where $N_{\text{text}}$ and $N_{\text{vision}}$ are the numbers of correct predictions by the textual and visual modalities, respectively.

**Visual Encoder Adaptation** The proposed Cross-modal Collaborative Pseudo-labeling strategy depends critically on the quality of visual prototypes. However, significant domain gaps persist between the CLIP and target downstream tasks, necessitating adaptation of the visual encoder.

To facilitate efficient adaptation of CLIP's visual encoder to downstream tasks, Visual Prompt Tuning (VPT) is commonly employed (Jia et al. 2022). However, the study (Wang et al. 2022) has found that CLIP exhibits a preference for certain categories during prediction, demonstrating an endogenous confirmation bias. This bias propagates into the computations of Eq.(7) and Eq.(9), affecting the quality of the generated pseudo-labels. Furthermore, low-quality pseudo-labels interfere with the calculation of $\mathcal{L}_{\text{pos}}$, thereby reducing the effectiveness of clean sample selection. Inspired by the idea of logit adjustment (Menon et al. 2020), we propose a debiased cross-entropy loss designed to mitigate the impact of CLIP's endogenous confirmation bias.

First, during each epoch, calculate the number of samples from $\mathcal{D}_{\text{NL}}$ and $\mathcal{D}_{\text{CL}}$ assigned to class $c$ by the model:

$$n_c = \sum_{i=1}^{|\mathcal{D}_{\text{NL}}|} \mathbb{I}\left(\hat{y}_i = c, \ p_i^{\text{M}}(c) \ge \tau^*\right) + \sum_{i=1}^{|\mathcal{D}_{\text{CL}}|} \mathbb{I}\left(y_i = c\right), \quad (11)$$

where $\tau^*$ is a pre-defined confidence threshold.

Then, calculate the model's preference degree for class $c$:

$$d_c = \frac{n_c}{\max_k n_k}. \tag{12}$$

Additionally, since CLIP tends to confuse categories with high similarity (Li et al. 2025), it is necessary to take this bias into account. The inter-class similarity is defined as:

$$\rho_{i,j} = \alpha \cdot \mathrm{sim}(\boldsymbol{T}_i^+, \boldsymbol{T}_j^+) + (1-\alpha) \cdot \mathrm{sim}(\boldsymbol{V}_i, \boldsymbol{V}_j), \tag{13}$$

where the $\alpha$ shares the same meaning and value as in Eq.(9).

Therefore, the logit adjustment term is defined as:

$$b_{i,j} = (1-d_j) \cdot \rho_{i,j} \cdot \mathbb{I}(i \neq j). \tag{14}$$

Finally, the debiased cross-entropy loss $\mathcal{L}_{\text{cls}}$ is defined as:

$$\mathcal{L}_{\text{cls}} = \frac{1}{|\mathcal{D}_{\text{CL}}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{CL}}} -\log \frac{e^{z_i^{y_i}}}{\sum_{j=1}^{C} e^{z_i^j + \gamma b_{y_i,j}}}, \tag{15}$$

where $z_i^j$ denotes the logit output for class $j$ given input $\mathbf{x}_i$, and $\gamma$ controls the strength of the logit adjustment term.

**Training Objective** The overall loss used in Phase 1 is formulated as follows:

$$\mathcal{L}_1 = \mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{neg}} + \lambda_1 \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{cls}}, \tag{16}$$

where $\lambda_1$ balances the strength of diversity regularization.

### 3.3 Phase 2: Noisy Label Refinement

After Phase 1, CLIP develops robust representation capabilities that demonstrate strong noise resilience. However, when confronting significant noise rates, leveraging the informational value of noisy-labeled samples becomes essential for performance enhancement. To maximize utilization of $\mathcal{D}_{\text{NL}}$, we implement a dedicated label refinement strategy in Phase 2. This generates additional high-quality supervision signals for final model adaptation in Phase 3.

**Dual-modal Prediction Consistency** The study (Menghini, Delworth, and Bach 2023) indicates that pseudo-labeling schemes with a confidence threshold ($p(y|x) > \tau$) to select samples to pseudo-label are ineffective for CLIP. Instead of relying on these, we propose a novel noisy label refinement strategy based on dual-modal prediction consistency.

Specifically, for each sample $(\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{NL}}$, we compute the category probability distributions from visual and textual modalities, $\boldsymbol{P}_i^{\text{V}}$ and $\boldsymbol{P}_i^{\text{L}}$, and obtain the predicted categories and their corresponding highest confidences:

$$\begin{aligned} \hat{y}_i^{\text{V}} &= \arg\max_c (p_i^{\text{V}}(c)), \quad p_i^{\text{V}} = \max_c p_i^{\text{V}}(c); \\ \hat{y}_i^{\text{L}} &= \arg\max_c (p_i^{\text{L}}(c)), \quad p_i^{\text{L}} = \max_c p_i^{\text{L}}(c). \end{aligned} \tag{17}$$

To quantify the reliability of predictions, we calculate the confidence margin for both modalities:

$$\Delta_i^{\text{V}} = p_i^{\text{V}} - \max_{k \neq \hat{y}_i^{\text{V}}} p_i^{\text{V}}(k), \quad \Delta_i^{\text{L}} = p_i^{\text{L}} - \max_{k \neq \hat{y}_i^{\text{L}}} p_i^{\text{L}}(k). \tag{18}$$

Similar to Eq.(17) and (18), the associated values of the fused probability distribution $\boldsymbol{P}_i^{\text{M}}$ in Eq.(9) can be obtained: $\hat{y}_i^{\text{M}}$, $p_i^{\text{M}}$, and $\Delta_i^{\text{M}}$. Given the significant disparity in recognition capabilities between the two modalities, we cannot treat them as equally important. Therefore, the textual modality is considered dominant when $\alpha > 0.5$; otherwise, the visual modality takes precedence. For simplicity in the following discussion, the associated values of the dominant modality are denoted as $\hat{y}_i^{\text{D}}$, $p_i^{\text{D}}$, and $\Delta_i^{\text{D}}$.

To improve the flexibility of sample filtering based on confidence margin, inspired by the idea of class-specific adaptive thresholds (Zhang et al. 2021), we assign an adaptive confidence margin threshold $\tau_k$ for each class, determined based on the class frequency within $\mathcal{D}_{\text{CL}}$. Let $f_{\max}$ be the highest class frequency and $f_k$ be the frequency of class $k$, then the adaptive threshold is defined as:

$$\tau_k = \max\left(\tau_{\text{base}} - \beta \cdot \left(1 - \frac{f_k}{f_{\max}}\right),\ \tau_{\min}\right), \tag{19}$$

where $\tau_{\text{base}}$ is the base threshold, $\beta$ controls the adjustment scale, $\tau_{\min}$ is the lower bound of the confidence margin.

Therefore, a sample is selected as a refined sample only if it satisfies $\hat{y}_i^{\text{D}} = \hat{y}_i^{\text{M}}$, and both $\Delta_i^{\text{D}} \geq \tau_{\hat{y}_i^{\text{D}}}$ and $\Delta_i^{\text{M}} \geq \tau_{\hat{y}_i^{\text{M}}}$.

**Calibrated Dual-Modal Refinement** For the refined samples selected through the above criteria, we treat the fused probability distribution $\boldsymbol{P}_i^{\text{M}}$ as soft label for the pseudo-label, instead of directly using hard label $\hat{y}_i^{\text{M}}$. Therefore, we can combine them to form the corrected sample subset $\mathcal{D}_{\text{PL}}$, which will be used with $\mathcal{D}_{\text{CL}}$ for fine-tuning in Phase 3.

### 3.4 Phase 3: Model Adaptation

Previous research (Wei et al. 2024) has shown that full fine-tuning (FFT) generally outperforms parameter-efficient fine-tuning (PEFT) on clean datasets. Based on this, in Phase 3, we use $\mathcal{D}_{\text{CL}}$ and $\mathcal{D}_{\text{PL}}$ to learn a linear classifier and perform FFT on a pre-trained vision model (e.g., ViT or ResNet).

To fully leverage the reliable supervision from clean samples, we train on $\mathcal{D}_{\text{CL}}$ using standard cross-entropy loss $\mathcal{L}_{\text{ce}}$.

Since the label accuracy in $\mathcal{D}_{\text{PL}}$ is generally lower than in $\mathcal{D}_{\text{CL}}$, to reduce the potential misguidance from incorrect pseudo-labels, we train only on $\mathcal{D}_{\text{CL}}$ during the warm-up stage. Once the model gains sufficient recognition ability on downstream tasks, $\mathcal{D}_{\text{PL}}$ is introduced for further adaptation.

To provide more stable guidance during model adaptation, we apply bidirectional KL-divergence loss $\mathcal{L}_{\text{kl}}$ on $\mathcal{D}_{\text{PL}}$ :

$$\mathcal{L}_{\text{kl}} = \frac{1}{|\mathcal{D}_{\text{PL}}|} \sum_{i=1}^{|\mathcal{D}_{\text{PL}}|} \frac{\mathrm{KL}(\boldsymbol{P}_i^{\text{M}} \parallel \boldsymbol{Q}_i) + \mathrm{KL}(\boldsymbol{Q}_i \parallel \boldsymbol{P}_i^{\text{M}})}{2}, \tag{20}$$

where $\boldsymbol{P}_i^{\text{M}}$ denotes the soft label, $\boldsymbol{Q}_i$ is the model's current predictive distribution, and $\mathrm{KL}(\cdot\|\cdot)$ is the KL-divergence.

Finally, the overall loss in Phase 3 is formulated as:

$$\mathcal{L}_2 = \mathcal{L}_{\text{ce}} + \lambda_2 \mathcal{L}_{\text{kl}}, \tag{21}$$

where $\lambda_2$ balances the strength of pseudo-label supervision.

## 4 Experiments

### 4.1 Experimental Settings

**Datasets** We first evaluate methods with synthetic datasets containing two types of controlled label noise: (1) **Symmetric noise** (Tanaka et al. 2018) randomly flips true label

| Method | *Sym.* 0.2 | *Sym.* 0.4 | *Sym.* 0.6 | *Ins.* 0.2 | *Ins.* 0.3 | *Ins.* 0.4 | *Sym.* 0.2 | *Sym.* 0.4 | *Sym.* 0.6 | *Ins.* 0.2 | *Ins.* 0.3 | *Ins.* 0.4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CUB-200-2011 | | | | | | Tiny-ImageNet | | | | | |
| CE | 80.76 | 73.09 | 55.42 | 80.36 | 75.80 | 69.62 | 81.77 | 76.53 | 73.17 | 80.75 | 78.83 | 74.80 |
| SCE | 82.81 | 78.12 | 63.31 | 81.91 | 78.31 | 71.25 | 79.23 | 76.24 | 71.76 | 78.96 | 77.80 | 74.47 |
| ELR | 77.70 | 68.26 | 50.17 | 78.32 | 73.16 | 63.57 | 79.40 | 77.13 | 73.74 | 79.98 | 77.13 | 73.74 |
| GMM | 75.79 | 64.39 | 42.84 | 75.73 | 69.95 | 56.13 | 81.91 | 80.37 | 43.47 | 81.84 | 81.26 | 79.01 |
| CLIPCleaner | 77.53 | 75.11 | 68.95 | 78.96 | 77.06 | 74.80 | 80.13 | 78.80 | 76.07 | 81.61 | 80.14 | 79.38 |
| GSM | 79.55 | 77.63 | 71.55 | 80.56 | 79.15 | 77.47 | 81.95 | 79.84 | 76.30 | 81.79 | 81.49 | 79.67 |
| DEFT | 83.05 | 79.24 | 73.08 | 82.53 | 81.39 | 79.34 | 82.91 | 82.48 | 80.60 | 83.37 | 82.69 | 80.52 |
| **DKAF (Ours)** | **83.33**$_{\pm 0.14}$ | **81.87**$_{\pm 0.22}$ | **78.22**$_{\pm 0.38}$ | **83.59**$_{\pm 0.13}$ | **82.98**$_{\pm 0.35}$ | **81.58**$_{\pm 0.32}$ | **83.17**$_{\pm 0.09}$ | **82.86**$_{\pm 0.23}$ | **81.88**$_{\pm 0.39}$ | **83.49**$_{\pm 0.06}$ | **82.98**$_{\pm 0.20}$ | **81.93**$_{\pm 0.26}$ |
| | Caltech-101 | | | | | | CIFAR-100 | | | | | |
| CE | 96.43 | 92.82 | 85.76 | 94.77 | 89.95 | 81.14 | 86.71 | 84.06 | 81.05 | 87.30 | 84.60 | 78.41 |
| SCE | 96.69 | 93.03 | 82.66 | 95.27 | 90.10 | 79.43 | 86.82 | 83.84 | 78.90 | 86.61 | 83.99 | 80.06 |
| ELR | 95.94 | 92.77 | 89.99 | 95.11 | 94.96 | 93.52 | 86.53 | 83.66 | 78.34 | 86.61 | 85.89 | 85.78 |
| GMM | 97.13 | 95.38 | 84.34 | 94.70 | 94.54 | 91.52 | 88.49 | 87.21 | 85.22 | 88.44 | 87.95 | 82.14 |
| CLIPCleaner | 94.04 | 93.93 | 93.36 | 94.23 | 93.97 | 93.59 | 87.93 | 85.51 | 82.97 | 88.43 | 86.99 | 84.98 |
| GSM | 97.85 | 96.88 | 94.81 | 97.36 | 96.33 | 95.11 | 88.21 | 86.67 | 83.27 | 88.81 | 87.51 | 85.08 |
| DEFT | 98.12 | 97.31 | 96.29 | 98.03 | 96.89 | 94.76 | 89.38 | 88.17 | 85.81 | 89.38 | 88.68 | 85.75 |
| **DKAF (Ours)** | **98.58**$_{\pm 0.12}$ | **98.43**$_{\pm 0.14}$ | **98.34**$_{\pm 0.24}$ | **98.55**$_{\pm 0.22}$ | **98.30**$_{\pm 0.28}$ | **97.52**$_{\pm 0.49}$ | **89.54**$_{\pm 0.05}$ | **88.82**$_{\pm 0.09}$ | **87.85**$_{\pm 0.26}$ | **89.56**$_{\pm 0.08}$ | **89.13**$_{\pm 0.24}$ | **87.71**$_{\pm 0.35}$ |
| | OxfordPets | | | | | | Stanford-Cars | | | | | |
| CE | 90.11 | 84.34 | 74.01 | 89.36 | 83.73 | 76.13 | 89.75 | 85.10 | 71.70 | 89.13 | 85.94 | 80.59 |
| SCE | 88.92 | 82.10 | 68.04 | 87.80 | 80.99 | 70.73 | 91.11 | 87.73 | 79.09 | 90.34 | 87.35 | 83.50 |
| ELR | 89.86 | 83.02 | 78.34 | 88.83 | 86.89 | 80.63 | 86.61 | 76.98 | 61.58 | 84.40 | 83.11 | 75.97 |
| GMM | 93.08 | 92.28 | 86.71 | 92.95 | 91.36 | 86.26 | 90.10 | 83.14 | 56.90 | 88.15 | 85.39 | 78.76 |
| CLIPCleaner | 90.93 | 89.85 | 87.12 | 91.13 | 90.86 | 87.98 | 90.44 | 87.84 | 84.63 | 90.06 | 89.37 | 88.13 |
| GSM | 91.99 | 91.15 | 88.21 | 91.97 | 90.87 | 89.82 | 91.22 | 89.76 | 86.35 | 91.57 | 90.71 | 89.88 |
| DEFT | 93.38 | 92.37 | 88.67 | 93.22 | 91.48 | 88.83 | 92.13 | 90.75 | 85.72 | 92.19 | 90.77 | 89.74 |
| **DKAF (Ours)** | **94.01**$_{\pm 0.13}$ | **93.54**$_{\pm 0.20}$ | **92.72**$_{\pm 0.21}$ | **94.25**$_{\pm 0.06}$ | **93.71**$_{\pm 0.04}$ | **92.90**$_{\pm 0.45}$ | **92.48**$_{\pm 0.12}$ | **91.59**$_{\pm 0.15}$ | **89.61**$_{\pm 0.28}$ | **92.38**$_{\pm 0.11}$ | **91.93**$_{\pm 0.23}$ | **91.49**$_{\pm 0.36}$ |

Table 2: Comparison of image classification accuracy (%) of different noisy label learning methods under various noise settings.

| Method | CIFAR-100N | Clothing-1M | WebVision | Avg. |
|---|---|---|---|---|
| CE | 72.41 | 69.75 | 84.64 | 75.60 |
| SCE | 72.52 | 70.49 | 82.88 | 75.30 |
| ELR | 72.83 | 72.14 | 79.32 | 74.76 |
| GMM | 76.06 | 70.03 | 84.88 | 76.99 |
| CLIPCleaner | 77.29 | 71.22 | 84.90 | 77.80 |
| GSM | 77.55 | 70.22 | 85.24 | 77.67 |
| DEFT | 79.04 | 72.44 | 85.12 | 78.87 |
| **DKAF (Ours)** | **81.91** | **73.36** | **85.35** | **80.21** |

Table 3: Comparison of classification accuracy (%) of different noisy label learning methods on real-world datasets.

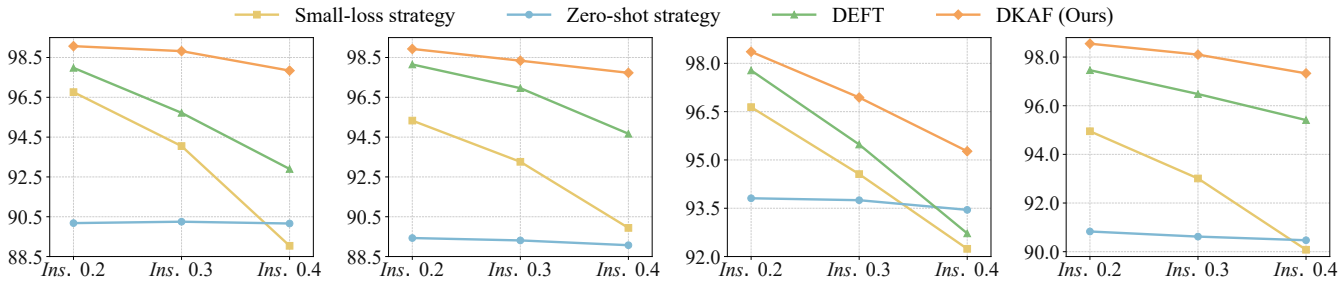
to a different class with a fixed probability; (2) **Instance-dependent noise** (Xia et al. 2020) simulates a more realistic scenario, where the corruption probability depends on both the true label and the image content. These two types of noise are introduced into six widely-used image classification datasets, including **CUB-200-2011** (Wah et al. 2011), **Tiny-ImageNet** (Le and Yang 2015), **Caltech-101** (Fei-Fei, Fergus, and Perona 2004), **CIFAR-100** (Krizhevsky, Hinton et al. 2009), **OxfordPets** (Parkhi et al. 2012), and **Stanford-Cars** (Krause et al. 2013). The noise rate is set to $r \in \{0.2, 0.4, 0.6\}$ for symmetric noise, and to $r \in \{0.2, 0.3, 0.4\}$ for instance-dependent noise. Following the settings of DEFT (Wei et al. 2024), we further evaluate DKAF on three real-world datasets: **CIFAR-100N** (Wei et al. 2021), **Clothing-1M** (Xiao et al. 2015), and **WebVision** (Li et al. 2017), which contain real-world noise.

**Implementation Details** We aim to improve CLIP's ability in clean sample selection and noisy label refinement. We use ViT-B/16 as the visual encoder and a jointly trained Transformer as the text encoder, fine-tuned via VPT and CoOp, respectively. We employ mini-batch SGD with a momentum of 0.9, a weight decay of $5 \times 10^{-4}$, and adopt the learning rate from DEFT (Wei et al. 2024). During training, random cropping and horizontal flipping are applied. For Phase 1, CLIP is fine-tuned for 10 epochs. During the warm-up stage, all pseudo-labels $\hat{y}_i$ used for computing $\mathcal{L}_{\text{pos}}$ are set to the labels $y_i$. Meanwhile, the SCE (Wang et al. 2019) is applied to all samples for computing $\mathcal{L}_{\text{cls}}$. Phase 3 is trained for 10 epochs on ViT-B/16 and includes a warm-up stage. All experiments were conducted on a single NVIDIA L20, with results averaged over three independent runs. Due to the large number of out-of-distribution (OOD) samples in WebVision and Clothing-1M, Phase 2 is not applied to both.

## 4.2 Main Results

**Evaluation Metric** Following the settings of prior related works (Wei et al. 2024), we report the best Top-1 test accuracy (%) achieved across all training iterations on downstream image classification tasks as the evaluation metric.

**Baselines** We compare our method against seven representative baselines under various noise types and levels on synthetic datasets. These include four classical approaches: CE (Cross-Entropy), SCE (Wang et al. 2019), ELR (Liu et al. 2020), and GMM (Li, Socher, and Hoi 2020); and three recent multimodal approaches: CLIPCleaner (Feng,




(a) OxfordPets (b) CIFAR-100 (c) Caltech-101 (d) Tiny-ImageNet

Figure 3: Comparison of F1-scores (%) achieved by different clean sample selection strategies under various noise settings.

Tzimiropoulos, and Patras 2024), GSM (Liang et al. 2025), and DEFT (Wei et al. 2024). In addition, we conduct comparisons on real-world datasets to further validate the generalization ability of our method. All methods leverage full fine-tuning to enhance the adaptation of the pre-trained vision model on downstream image classfication tasks.

**Results Analysis** As shown in Tab.2, our method consistently achieves the best results across all six datasets and various noise settings, with particularly strong performance under high-noise conditions. For instance, under the challenging 60% symmetric noise setting, our method surpasses the second-best method by 3.26%, 4.05%, and 5.14% on Stanford-Cars, OxfordPets, and CUB-200-2011, respectively. Moreover, our method also achieves superior results on real-world datasets (Tab.3), outperforming all baselines and demonstrating strong robustness in practical scenarios.

### 4.3 Performance for Clean Sample Selection

**Evaluation Metrics** We use precision, recall, and F1-score to assess the performance of clean sample selection. Precision measures the ratio of truly clean samples within $\mathcal{D}_{\text{CL}}$, while recall reflects the ratio of all clean samples in $\mathcal{D}$ that are successfully identified. F1-score is the harmonic mean of precision and recall, offering a balanced assessment. Therefore, we report it as the main evaluation metric.

**Baselines** We compare our method with three common strategies for identifying clean samples under noisy label settings: (1) Small-loss strategy (Jiang et al. 2018) identifies clean samples based on their low training loss, which is also employed by the classic method GMM (Li, Socher, and Hoi 2020). (2) Zero-shot strategy (Radford et al. 2021) leverages CLIP's zero-shot predictions to detect noisy labels, with two main implementations: label consistency (Feng, Tzimiropoulos, and Patras 2024) and confidence thresholding (Liang et al. 2025). We report the best result among these two approaches. (3) DEFT (Wei et al. 2024) is a vanilla self-training method based on CLIP and achieves strong performance in clean sample selection.

**Results Analysis** As shown in Fig.3, the compared methods behave differently under various noise settings: the small-loss strategy is highly sensitive to noise levels, with performance dropping sharply as the noise rate increases; the zero-shot strategy remains relatively stable across various noise rates, but its overall effectiveness is limited due to the lack of adaptation to specific downstream tasks; DEFT generally performs more robustly, but shows limited performance under high-noise conditions on certain datasets. In contrast, our method consistently outperforms all baselines, achieving the highest F1-scores across different noise scenarios, and demonstrating strong adaptability. Notably, under the 40% noise rate setting on Caltech-101, both DEFT and the small-loss strategy yield lower F1-scores than the zero-shot strategy, indicating limited robustness in high-noise conditions. In contrast, our method still maintains the leading position, showing its strong noise resistance.

### 4.4 Ablation Study

To evaluate the effectiveness of each phase in DKAF, we conducted an ablation study by progressively removing each phase and assessing the model's image classification accuracy. As shown in Tab.4, incorporating Phase 1, Phase 2, and Phase 3 together yields the highest classification accuracy across all datasets, indicating that each phase makes a significant contribution to improving the overall performance.

| Method | P1 | P2 | P3 | CIFAR-100 | Cars | CUB | Avg. |
|---|---|---|---|---|---|---|---|
| **DKAF (Ours)** | ✓ | ✓ | ✓ | **87.85** | **89.61** | **78.22** | **85.23** |
| | ✓ | - | ✓ | 86.41 | 86.94 | 74.35 | 82.57 |
| | ✓ | - | - | 85.20 | 77.04 | 63.91 | 75.38 |
| | - | - | - | 81.05 | 71.70 | 55.42 | 69.39 |

Table 4: Ablation results on three datasets under *Sym.* 0.6.

## 5 Conclusion

In this paper, we addressed the challenge of endogenous confirmation bias in noisy label learning for vision-language models. To tackle this issue, we proposed the Debiased Knowledge Adaptation Framework (DKAF), a multi-phase approach that integrates clean sample selection, noisy label refinement, and model adaptation. By explicitly mitigating the over-reliance on pretrained knowledge, DKAF facilitates more effective noisy label detection and refinement. Extensive experiments on nine benchmark datasets demonstrate the effectiveness and robustness of our method, underscoring its potential for real-world noisy label learning tasks.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (62306085), Shenzhen College Stability Support Plan (GXWD20231130151329002), CCF-ALIMAMA TECH Kangaroo Fund (CCF-ALIMAMA OF 2025001), Shenzhen Science and Technology Program (KQTD20240729102207002).

## References

Alabdulmohsin, I.; Wang, X.; Steiner, A.; Goyal, P.; D'Amour, A.; and Zhai, X. 2024. Clip the bias: How useful is balancing data in multimodal learning? *arXiv preprint arXiv:2403.04547*.

Bordes, F.; Pang, R. Y.; Ajay, A.; Li, A. C.; Bardes, A.; Petryk, S.; Mañas, O.; Lin, Z.; Mahmoud, A.; Jayaraman, B.; et al. 2024. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*.

Chen, G.; Yao, W.; Song, X.; Li, X.; Rao, Y.; and Zhang, K. 2022. Plot: Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253*.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *conference on computer vision and pattern recognition workshop*, 178–178. IEEE.

Feng, C.; Tzimiropoulos, G.; and Patras, I. 2024. Clipcleaner: Cleaning noisy labels with clip. In *Proceedings of the 32nd ACM international conference on multimedia*, 876–885.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2024. Clip-adapter: Better vision-language models with feature adapters. *International journal of computer vision*, 132(2): 581–595.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31: 8527–8537.

Huang, T.; Chu, J.; and Wei, F. 2022. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European conference on computer vision*, 709–727. Springer.

Jiang, L.; Zhou, Z.; Leung, T.; Li, L.-J.; and Fei-Fei, L. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, 2304–2313. PMLR.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19113–19122.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, 554–561.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, J.; Socher, R.; and Hoi, S. C. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*.

Li, S.; Liu, F.; Hao, Z.; Wang, X.; Li, L.; Liu, X.; Chen, P.; and Ma, W. 2025. Logits deconfusion with CLIP for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 25411–25421.

Li, W.; Wang, L.; Li, W.; Agustsson, E.; and Van Gool, L. 2017. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*.

Liang, C.; Zhu, L.; Shi, H.; and Yang, Y. 2025. Combating label noise with a general surrogate model for sample selection. *International Journal of Computer Vision*, 133(6): 3166–3179.

Liu, S.; Niles-Weed, J.; Razavian, N.; and Fernandez-Granda, C. 2020. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33: 20331–20342.

Menghini, C.; Delworth, A.; and Bach, S. 2023. Enhancing clip with clip: Exploring pseudolabeling for limited-label prompt tuning. *Advances in neural information processing systems*, 36: 60984–61007.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2020. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *conference on computer vision and pattern recognition*, 3498–3505. IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Reed, S.; Lee, H.; Anguelov, D.; Szegedy, C.; Erhan, D.; and Rabinovich, A. 2014. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*.

Shen, Y.; and Sanghavi, S. 2019. Learning with bad training data via iterative trimmed loss minimization. In *International conference on machine learning*, 5739–5748. PMLR.

Song, H.; Kim, M.; Park, D.; Shin, Y.; and Lee, J.-G. 2022. Learning from noisy labels with deep neural networks: A

survey. *IEEE transactions on neural networks and learning systems*, 34(11): 8135–8153.

Tanaka, D.; Ikami, D.; Yamasaki, T.; and Aizawa, K. 2018. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5552–5560.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset. Technical report, California Institute of Technology.

Wang, X.; Wu, Z.; Lian, L.; and Yu, S. X. 2022. Debiased learning from naturally imbalanced pseudo-labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14647–14657.

Wang, Y.; Ma, X.; Chen, Z.; Luo, Y.; Yi, J.; and Bailey, J. 2019. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, 322–330.

Wei, H.; Feng, L.; Chen, X.; and An, B. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13726–13735.

Wei, J.; Zhu, Z.; Cheng, H.; Liu, T.; Niu, G.; and Liu, Y. 2021. Learning with noisy labels revisited: A study using real-world human annotations. *arXiv preprint arXiv:2110.12088*.

Wei, T.; Li, H.-T.; Li, C.; Shi, J.-X.; Li, Y.-F.; and Zhang, M.-L. 2024. Vision-language models are strong noisy label detectors. *Advances in neural information processing systems*, 37: 58154–58173.

Xia, X.; Liu, T.; Han, B.; Wang, N.; Gong, M.; Liu, H.; Niu, G.; Tao, D.; and Sugiyama, M. 2020. Part-dependent label noise: Towards instance-dependent label noise. *Advances in neural information processing systems*, 33: 7597–7610.

Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; and Wang, X. 2015. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2691–2699.

Xing, Y.; Kang, J.; Xiao, A.; Nie, J.; Shao, L.; and Lu, S. 2023. Rewrite caption semantics: Bridging semantic gaps for language-supervised semantic segmentation. *Advances in neural information processing systems*, 36: 68798–68809.

Xu, C.; Tao, D.; and Xu, C. 2013. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*.

Yi, K.; and Wu, J. 2019. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7017–7025.

Zhang, B.; Wang, Y.; Hou, W.; Wu, H.; Wang, J.; Okumura, M.; and Shinozaki, T. 2021. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in neural information processing systems*, 34: 18408–18419.

Zhang, Z.; and Sabuncu, M. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31: 8778–8788.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9): 2337–2348.

Zhou, Z.-H. 2018. A brief introduction to weakly supervised learning. *National science review*, 5(1): 44–53.