Dynamic Positive CLIP Distillation for Multi-Labeled Complementary Label Learning

Anonymous authors
Paper under double-blind review

Abstract

Multi-labeled complementary label learning (MLCLL) trains multi-label classifiers from complementary labels, which indicate classes absent from a sample rather than its positive labels. Since verifying absent classes can be easier than exhaustively annotating all positive labels, MLCLL provides a cost-effective solution for multi-label learning (MLL). Existing methods mainly learn from complementary labels, which reveal only a subset of negative classes and provide no positive supervision. Our analysis shows that such non-exhaustive complementary supervision may leave residual uncertainty about the positive-label set, while only negative labels may result in a degenerate solution that assigns low probabilities to all classes. Pretrained vision-language models offer a promising way to recover the missing positive semantics, but existing approaches typically require observed positive labels as semantic anchors and therefore cannot be directly applied to MLCLL. To address this limitation, we propose CoDi, a CLIP-based framework for MLCLL without positive labels. CoDi first employs a frozen CLIP teacher with positive-only prompts to construct positive evidence from cached global predictions and dynamically updated random-local and class-aware views guided by an exponential moving average (EMA) student. It then introduces reliability aware pseudo-positive learning to select, weight, and temporally stabilize pseudo-positive labels while preserving observed complementary labels as reliable negative supervision. The resulting positive knowledge is distilled into a standalone student model, enabling CLIP-free inference. Experiments on four benchmark datasets demonstrate the effectiveness of CoDi.

1 Introduction

Multi-label learning (MLL) aims to identify a set of positive labels associated with an unseen sample (Zhang & Zhou, 2014). As real-world samples are often associated with multiple semantic concepts, MLL has been widely applied to various domains, including medical diagnosis (Wang et al., 2017; Liao et al., 2026; Mahapatra et al., 2024), video understanding (Zhu et al., 2024; Li et al., 2025), and AI for Science (Kulmanov et al., 2024; Hu et al., 2024; Wang et al., 2025). However, collecting exhaustive multi-label annotations is costly, as annotators must examine a potentially large label space to determine all relevant labels, particularly when the label space is large or target objects are small or rare (Gao et al., 2023; 2024; 2025). To reduce this annotation burden, weakly supervised MLL has emerged as a practical alternative, which enables classifiers to be trained with incomplete rather than exhaustive label supervision (Tang et al., 2024; Lyu et al., 2025).

Multi-labeled complementary-label learning (MLCLL) offers a cost-effective weakly supervised paradigm in which each sample is annotated with complementary labels indicating classes that are absent (Gao et al., 2023; 2026b). Since verifying absent classes can be easier than exhaustively annotating all positive labels, MLCLL reduces annotation cost while providing reliable negative supervision. Existing MLCLL methods learn from the observed complementary labels without positive supervision, while such complementary supervision typically covers only a small subset of the negative classes (Gao et al., 2023; Zhu et al., 2025;

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.

Gao et al., 2026b). Since the remaining unannotated labels contain both positive labels and unrevealed negative labels, non-exhaustive complementary labels are insufficient to fully determine the underlying positive-label set. Our analysis shows that such non-exhaustive complementary supervision may leave residual uncertainty about the positive-label set, while negative labels alone may result in a degenerate solution that assigns low probabilities to all classes. These limitations motivate us to seek positive semantic evidence beyond the observed complementary annotations. Pretrained vision-language models offer a promising source of such missing positive semantics through their rich image-text representations (Xing et al., 2024; Rawlekar et al., 2025; Tran et al., 2025; Sun et al., 2026).

Pretrained vision-language models (VLMs) offer a promising way to supply such information. Models such as CLIP align images and natural-language descriptions in a shared representation space, which enables category-level recognition through textual class prompts (Radford et al., 2021). Recent studies have exploited such semantic knowledge to generate pseudo-positive supervision for single-positive MLL (Xing et al., 2024; Tran et al., 2025), while others improve vision-language distillation for unsupervised MLL by combining global predictions with localized visual evidence (Kim & Shim, 2025). However, these methods cannot be directly applied to MLCLL. Single-positive MLL methods rely on observed positive labels as reliable semantic anchors to calibrate CLIP predictions and discover additional positives, while MLCLL provides no observed positive labels. Unsupervised MLL methods instead derive positive supervision largely from CLIP predictions themselves, which may introduce noisy pseudo-labels in the absence of positive anchors. Moreover, neither line of work explicitly exploits the reliable negative constraints provided by complementary labels.

To address these challenges, we propose CoDi, multi-labeled Complementary-label learning via CLIP Distillation, a CLIP-based framework that recovers the missing positive semantics without requiring any observed positive labels. CoDi first employs a frozen CLIP teacher with positive-only prompts to construct positive evidence from cached global predictions and dynamically updated random-local and class-aware views guided by an exponential moving average (EMA) student. Based on this multi-view evidence, we introduce reliability-aware pseudo-positive learning to select and weight reliable positive candidates and temporally stabilize them during training, while preserving the observed complementary labels as trustworthy negative supervision. The recovered positive knowledge is then distilled into the student model, enabling the complementary negative supervision and CLIP-derived positive semantics to be jointly exploited. After training, the CLIP teacher is discarded and only the standalone student is retained for inference, introducing no additional CLIP-related inference cost. Experiments on four benchmark datasets demonstrate the effectiveness of CoDi. Our main contributions are summarized as follows:

1) To the best of our knowledge, we present the first VLM-based framework for MLCLL. We further analyze the supervision limitations of MLCLL, showing that non-exhaustive complementary labels may leave residual uncertainty about the positive-label set, while negative-only learning can lead to degenerate solutions.

2) We propose CoDi, which constructs positive evidence from a frozen positive-prompt CLIP teacher using global, random-local, and EMA-student-guided class-aware views, thereby providing positive semantics beyond the observed negative supervision.

3) We develop reliability-aware pseudo-positive learning to select, weight, and temporally stabilize CLIP derived positive candidates while explicitly preserving observed complementary labels as reliable negatives. The recovered positive knowledge is distilled into a standalone student, enabling effective CLIP-free inference.

The remainder of this paper is structured as follows. Section 2 reviews the related work. Section 3 introduces the proposed method. Section 4 presents the experimental results, and Section 5 concludes the paper.

2 Related Work

In MLCLL, each sample is only associated with one or more complementary labels. Gao et al. (2023) first formulated MLCLL and derived an unbiased risk estimator with theo-

retical convergence guarantees. Subsequent work introduced a correlation-aware correction method that exploits label dependencies to estimate the transition matrix (Gao et al., 2024) and extended MLCLL to accommodate multiple complementary labels per sample (Gao et al., 2025). Beyond risk estimation, ComRank (Zhu et al., 2025) adopts a ranking objective that places complementary labels below non-complementary labels. More recent studies improve robustness to noisy supervision through tailored mixup, label-correlation modeling, and consistency regularization (Gao et al., 2026b), and incorporate hard logical constraints among labels as additional weak supervision (Gao et al., 2026a). Existing MLCLL methods primarily derive supervision from complementary labels and inter-label relationships, leaving the visual-semantic knowledge encoded in pretrained vision-language models unexplored. Without observed positive labels, these methods lack positive semantic evidence for identifying positive labels. Our approach addresses this gap by investigating how a pretrained CLIP teacher can supply the missing positive semantics while preserving complementary labels as reliable negative supervision.

Large-scale VLMs such as CLIP (Radford et al., 2021) provide transferable visual-semantic knowledge that reduces annotation requirements in MLL. DualCoOp (Sun et al., 2022) and DualCoOp++ (Hu et al., 2023) adapt CLIP through positive–negative prompts and evidence-guided context optimization, respectively, while PositiveCoOp (Rawlekar et al., 2025) favors positive prompts given the limited reliability of negative prompts. Other approaches address incomplete annotations using structured semantic priors (Ding et al., 2023) or dense vision-language relationships (Chen et al., 2026a). For single-positive MLL, VLPL (Xing et al., 2024) generates CLIP-based pseudo-labels, while AEVLP (Tran et al., 2025) improves their reliability through multi-focus generation and noise-robust learning. For annotation-free learning, CDUL (Abdelfattah et al., 2023) combines global and local CLIP predictions, and CCD (Kim & Shim, 2025) addresses view sensitivity and prediction bias through classifier-guided local views and calibrated distillation. Further work reduces VLMs' bias toward salient objects to improve recognition coverage (Chen et al., 2026b). These methods primarily target settings with positive annotations or without label constraints. In contrast, MLCLL provides reliable negative constraints but no observed positive labels. CoDi combines these constraints with positive semantics recovered from a frozen VLM and distills this knowledge into a standalone multi-label classifier.

3 Methodology

In this section, we first formulate the MLCLL problem and theoretically analyze the insufficiency of non-exhaustive complementary supervision. We then present CoDi, a CLIP-based framework that complements reliable negative supervision with positive semantic evidence without observed positive labels. As illustrated in Figure 1, CoDi mainly consists of two components: multi-view positive evidence construction and reliability-aware pseudo-positive learning. The former combines class-wise calibrated CLIP predictions across multiple views, while the latter selects reliable pseudo-positive targets under the guidance of observed complementary labels and maintains them in a bank. Finally, we describe joint optimization and training using complementary labels and recovered positive supervision, with only the EMA student retained for CLIP-free inference.

3.1 Preliminaries

Let $\mathcal{X}$ denote the input space and $\mathcal{Y} = \{0, 1\}^C$ the label space with C ($C > 2$) labels. In MLL, let $D = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$ denote a dataset of N samples, where $\boldsymbol{x}_i \in \mathcal{X}$ and $\boldsymbol{y}_i \in \mathcal{Y}$ are the input and its corresponding binary label vector, respectively. Specifically, $y_{ic} = 1$ indicates that the c-th label is positive for $\boldsymbol{x}_i$, while $y_{ic} = 0$ indicates that it is negative.

Since positive labels are unavailable in MLCLL, each sample $\boldsymbol{x}_i$ is associated with a complementary label vector $\bar{\boldsymbol{y}}_i \in \{1, 0\}^C$, where $\bar{y}_{ic} = 1$ indicates that the c-th label is negative for $\boldsymbol{x}_i$, while $\bar{y}_{ic} = 0$ denotes an unobserved label whose true status may be either positive or negative. Notably, complementary labels are sampled only from the true negative labels of each sample. The goal of MLCLL is to learn a multi-label classifier $\boldsymbol{f} : \mathcal{X} \to [0, 1]^C$ that predicts the label set of an unseen sample by thresholding its class-wise prediction scores.

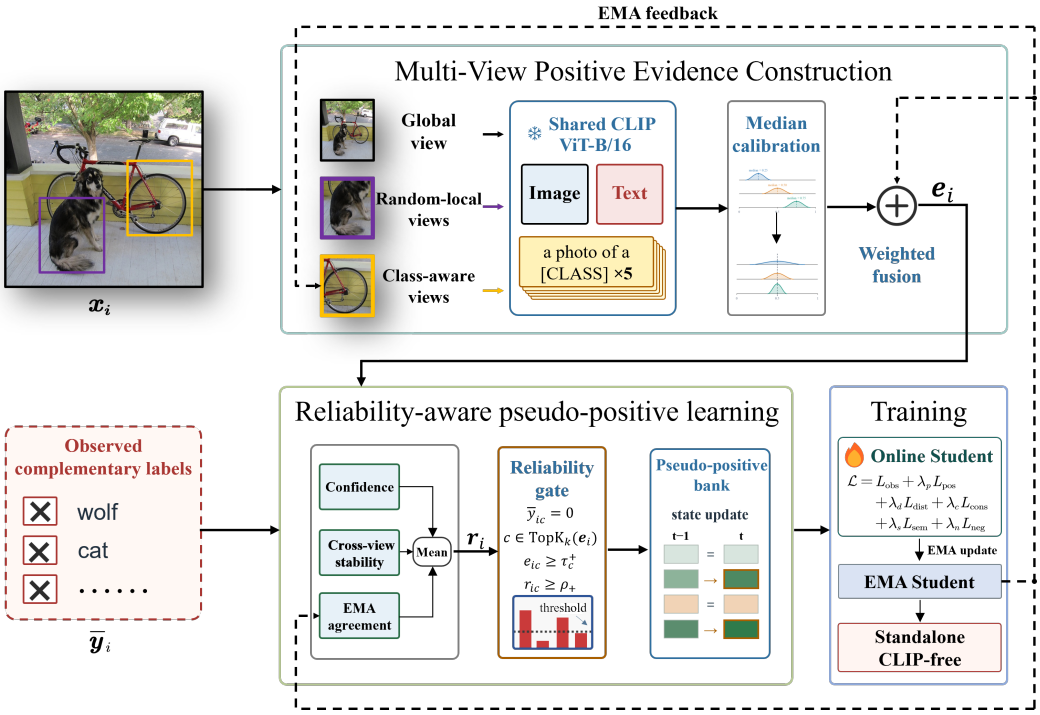


Figure 1: The framework of CoDi.

3.2 Information Insufficiency of Complementary Supervision

Existing MLCLL methods rely on observed complementary labels, which typically cover only a small subset of each sample's negative classes (Gao et al., 2023; 2026b), leaving unobserved labels potentially positive or negative. We therefore quantify the information gap between complementary and full supervision through conditional mutual information.

Theorem 1. Let $(X, Y, \bar{Y})$ denote the random variables corresponding to $(\boldsymbol{x}_i, \boldsymbol{y}_i, \bar{\boldsymbol{y}}_i)$, respectively. The information about Y provided by complete supervision and complementary supervision is measured by $I(Y; Y \mid X)$ and $I(Y; \bar{Y} \mid X)$, respectively. Their information gap satisfies $I(Y; Y \mid X) - I(Y; \bar{Y} \mid X) = H(Y \mid X, \bar{Y}) \geq 0$.

The proof is provided in Appendix A. Here, $H(Y \mid X, \bar{Y})$ measures the residual uncertainty about Y after observing the input and its complementary labels. A zero gap means that $(X, \bar{Y})$ uniquely determines Y, as in exhaustive complementary annotation where $\bar{\boldsymbol{y}}_i = \mathbf{1} - \boldsymbol{y}_i$. A positive gap indicates that complementary supervision is strictly less informative than complete supervision.

Beyond the information insufficiency discussed above, complementary-only supervision also poses an optimization-level limitation. Since observed complementary labels provide only negative constraints, they encourage the corresponding class probabilities to decrease without requiring any positive predictions, as formalized below.

Proposition 1. Let $O^- = \{(\boldsymbol{x}_i, c) \mid \bar{y}_{ic} = 1\}$ denote the observed complementary label set, and consider the negative-only objective $\ell = \frac{1}{|O^-|} \sum_{(\boldsymbol{x}_i, c) \in O^-} l^-(f_c(\boldsymbol{x}_i))$, where $l^- : [0, 1] \to \mathbb{R}_+$ is a negative-label loss whose minimum is attained at $f_c(\boldsymbol{x}_i) = 0$. Then the all-zero prediction $\boldsymbol{f}(\boldsymbol{x}_i) = \mathbf{0}$ is a global minimizer of ℓ.

The proof is provided in Appendix B. Proposition 1 shows that negative-only supervision admits a degenerate optimum in which all class probabilities collapse to zero, as no positive constraint prevents such suppression. Together with Theorem 1, this motivates incorporating additional positive semantic evidence while retaining the observed complementary labels as reliable negative supervision.

3.3 Multi-View Positive Evidence Construction

To supplement the available supervision, CoDi uses a frozen CLIP teacher to provide positive semantic evidence for training the student classifier $\boldsymbol{f}$. An EMA copy of the student, denoted by $\bar{\boldsymbol{f}}$, guides class-aware view construction and stabilizes training. Let E_{img} and E_{text} be the frozen CLIP image and text encoders, respectively, with ℓ_2-normalized outputs. For each class c, we construct a class prototype by averaging the embeddings of R positive prompts and normalizing the result:

$$\boldsymbol{t}_c = \text{norm}\left(\frac{1}{R}\sum_{r=1}^{R} E_{\text{text}}(T_r(c))\right), \tag{1}$$

where $T_r(c)$ is the r-th positive prompt for class c and $\text{norm}(\cdot)$ denotes ℓ_2-normalization. For a visual view v of image $\boldsymbol{x}_i$, the teacher computes a class-presence score:

$$s_{ic}^{(v)} = \sigma\left(4\exp(\ell_T)/100\,\langle E_{\text{img}}(v(\boldsymbol{x}_i)), \boldsymbol{t}_c\rangle\right), \tag{2}$$

where $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product, which is equivalent to cosine similarity for ℓ_2-normalized image and text features. The parameter ℓ_T denotes the pretrained CLIP logit-scale parameter in log space and is kept fixed during training.

All medians and quantiles are estimated from the N training samples used for model optimization. To align score scales across classes, we calibrate each class using its median global-view score:

$$q_{ic}^{(v)} = \text{clip}_{[0,1]}\left(w_{\text{wide}} s_{ic}^{(v)} / \max(\text{clip}_{[0.05,0.95]}(\text{median}_{j=1,\ldots,N}\, s_{jc}^{(g)}), \epsilon)\right), \tag{3}$$

where g denotes the full-image view, $w_{\text{wide}} > 0$ controls the calibration scale, $\text{clip}_{[a,b]}$ clips its input to $[a,b]$, and $\epsilon > 0$ ensures numerical stability. When clipping and numerical stabilization are inactive, this transformation maps the class-wise global median to w_{wide}, aligning score scales without positive annotations.

Global views may miss small concepts, we therefore sample two random local crops per image, then evaluate them independently with CLIP. Here, we aggregate their raw scores with a global fallback:

$$s_{ic}^{(r)} = \max\left\{s_{ic}^{(g)}, \frac{s_{ic}^{(r_1)} + s_{ic}^{(r_2)}}{2}\right\}. \tag{4}$$

Applying equation 3 yields the random-local evidence $q_{ic}^{(r)}$. The global fallback preserves full-image evidence when the crops miss relevant content.

To capture class-specific regions, we select candidate classes from the union of the top-ranked classes under the calibrated global CLIP scores and EMA predictions. We use class activation maps (CAMs) from $\bar{\boldsymbol{f}}$ to extract class-aware crops and evaluate them with CLIP. For each class, we take the maximum of its global and corresponding crop scores before calibration to obtain $q_{ic}^{(a)}$, retaining the global score when no valid crop is available. We combine these sources with the EMA predictions to obtain the final positive evidence:

$$e_{ic} = \alpha_g q_{ic}^{(g)} + \alpha_r q_{ic}^{(r)} + \alpha_a q_{ic}^{(a)} + \alpha_e \bar{f}_c(\boldsymbol{x}_i), \tag{5}$$

where $\bar{f}_c(\boldsymbol{x}_i)$ is the EMA probability for class c. The coefficients are nonnegative and sum to 1, ensuring $e_{ic} \in [0,1]$. During warm-up, only global and random-local CLIP evidence is used ($\alpha_a = \alpha_e = 0$). Joint training then incorporates class-aware evidence and EMA predictions, progressively enriching the frozen teacher's semantic priors with student-guided, class-specific evidence.

3.4 Reliability-aware Pseudo-positive Learning

Although multi-view aggregation broadens the coverage of positive semantics, CLIP predictions can still produce false positives. This is because no observed positive labels are available in MLCLL as semantic anchors. Therefore, we use observed complementary labels

as negative references to derive class-wise thresholds and assess candidate reliability before constructing pseudo-positive supervision.

For each class c, we define a negative anchor τ_c^- and a positive selection threshold τ_c^+:

$$\tau_c^- = \begin{cases} \min\left\{\tau_{\max}^-, \text{Quantile}_\eta\left(\{q_{ic}^{(g)} \mid \bar{y}_{ic} = 1\}\right)\right\}, & \sum_{i=1}^N \bar{y}_{ic} \geq n_{\min}, \\ \tau_{\max}^-, & \text{otherwise}, \end{cases} \tag{6}$$

$$\tau_c^+ = \max\left\{\tau_{\min}, \text{Quantile}_\zeta\left(\{q_{ic}^{(g)}\}_{i=1}^N\right), \tau_c^- + \delta\right\},$$

where η and ζ are quantile levels. Estimating the negative anchor requires at least $n_{\min}$ observed negatives, with $\tau_{\max}^-$ as its upper bound and fallback value. The positive threshold is lower-bounded by $\tau_{\min}$ and lies at least $\delta > 0$ above the anchor.

For each unobserved sample-class pair, we assess the reliability of its positive evidence through confidence, cross-view stability, and agreement with the EMA student:

$$r_{ic} = \frac{1}{3}\left[\text{clip}_{[0,1]}\left(\frac{e_{ic} - \tau_c^+}{\max(1 - \tau_c^+, \epsilon)}\right) + \exp\left(-\frac{\text{Var}(\mathcal{V}_{ic})}{\sigma_v^2}\right) + 1 - \left|e_{ic} - \bar{f}_c(\boldsymbol{x}_i)\right|\right], \tag{7}$$

where $\mathcal{V}_{ic}$ contains the calibrated branch scores, with global scores used as fallbacks, and $\sigma_v > 0$ controls sensitivity to cross-view variance. A new observation is accepted when it satisfies the following ranking, confidence, and reliability conditions:

$$A = \{(\boldsymbol{x}_i, c) \mid \bar{y}_{ic} = 0, c \in \text{TopK}_k(\boldsymbol{e}_i), e_{ic} \geq \tau_c^+, r_{ic} \geq \rho_+\}, \tag{8}$$

where $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{iC})$, $\text{TopK}_k(\cdot)$ returns the indices of the k largest entries, and ρ_+ is the reliability threshold. We maintain a pseudo-positive bank whose active sample-class pairs form O^+. Each entry stores a soft target $\hat{y}_{ic}$ and a reliability weight w_{ic}. Initially, all entries are inactive, with $\hat{y}_{ic} = 0.5$ and $w_{ic} = 0$. Each accepted observation updates the stored target and weight through EMAs:

$$\hat{y}_{ic} \leftarrow \beta\hat{y}_{ic} + (1 - \beta)e_{ic}, \quad w_{ic} \leftarrow \beta w_{ic} + (1 - \beta)r_{ic}, \tag{9}$$

where $\beta \in [0, 1)$ is the bank momentum. When no new observation is accepted for an entry, its target is retained and its weight decays as $w_{ic} \leftarrow \beta w_{ic}$. Entries with updated weights at or below a small cutoff are deactivated.

3.5 Joint Student Optimization and Inference

Using the complementary labels and recovered positive evidence, we optimize the student $\boldsymbol{f}$ with the following losses. For the observed complementary labels, the corresponding loss is defined as:

$$L_{\text{obs}} = -\frac{1}{N}\sum_{(\boldsymbol{x}_i, c) \in O^-} p_{ic}^{\gamma_-} \log(1 - p_{ic}), \text{ where } p_{ic} = \max(f_c(\boldsymbol{x}_i) - \xi, 0). \tag{10}$$

Here, the shift $\xi \in [0, 1)$ suppresses easy negatives, while $\gamma_- \geq 0$ controls the emphasis on hard negatives. Motivated by the generalized pseudo-label loss from Tran et al. (2025), we define a tailored loss function for the pseudo-positive supervision as:

$$L_{\text{pos}} = \frac{\sum_{(\boldsymbol{x}_i, c) \in O^+} w_{ic}[1 - (\hat{y}_{ic} f_c(\boldsymbol{x}_i) + (1 - \hat{y}_{ic})(1 - f_c(\boldsymbol{x}_i)))^q]/q}{\max(1, \sum_{(\boldsymbol{x}_i, c) \in O^+} w_{ic})}, \tag{11}$$

where $q \in (0, 1]$ controls the sensitivity to pseudo-positive noise. This objective supplies positive constraints absent from complementary labels, where w_{ic} downweights uncertain targets, while larger q reduces the gradient contribution of predictions with low agreement with the stored targets.

To transfer evidence beyond the selected pseudo-positive bank entries, we distill e_{ic} over all unobserved pairs by the following objective:

$$L_{\text{dist}} = \frac{1}{N(C-1)}\sum_{i=1}^N \sum_{c=1, \bar{y}_{ic} \neq 1}^C (1 - \bar{y}_{ic})(f_c(\boldsymbol{x}_i) - e_{ic})^2. \tag{12}$$

To improve robustness to augmentation, we align student predictions on a strong view $\boldsymbol{x}_i^{(s)}$ with EMA predictions on the weak view:

$$L_{\text{cons}} = -\frac{1}{NC}\sum_{i=1}^{N}\sum_{c=1}^{C}\left[\bar{f}_c(\boldsymbol{x}_i)\log f_c(\boldsymbol{x}_i^{(s)}) + (1-\bar{f}_c(\boldsymbol{x}_i))\log(1-f_c(\boldsymbol{x}_i^{(s)}))\right], \tag{13}$$

where predictions on $\boldsymbol{x}_i$ use the weak view. Let $\boldsymbol{z}_i$ be the normalized student feature projected into the text embedding space. We align it with the positive prototypes of the active pseudo-positive classes:

$$L_{\text{sem}} = \frac{1}{|O^+|}\sum_{(\boldsymbol{x}_i,c)\in O^+}(1-\boldsymbol{z}_i^\top \boldsymbol{t}_c). \tag{14}$$

For observed negative pairs, we encourage greater similarity to a learned unit negative prototype $\boldsymbol{n}_c$ than to $\boldsymbol{t}_c$:

$$L_{\text{neg}} = \frac{1}{\max(1,|O^-|)}\sum_{(\boldsymbol{x}_i,c)\in O^-}\left[m+\boldsymbol{z}_i^\top \boldsymbol{t}_c - \boldsymbol{z}_i^\top \boldsymbol{n}_c\right]_+, \tag{15}$$

where $[u]_+ = \max(0,u)$ and $m = 0.2$ is the margin. The prototype $\boldsymbol{n}_c$ lies in the same embedding space and is learned without negative text prompts. Finally, the student is jointly optimized by minimizing

$$\mathcal{L} = L_{\text{obs}} + \lambda_p L_{\text{pos}} + \lambda_d L_{\text{dist}} + \lambda_c L_{\text{cons}} + \lambda_s L_{\text{sem}} + \lambda_n L_{\text{neg}}, \tag{16}$$

where the nonnegative coefficients balance the loss terms. After warm-up and joint training, we freeze the bank and stop CLIP image evaluations for final fine-tuning. At inference, only the EMA student $\bar{\boldsymbol{f}}$ is retained, requiring a single full-image forward pass.

4 Experiments

4.1 Experimental Settings

We evaluate CoDi on PASCAL VOC 2007 (VOC07) (Everingham et al., 2010), MS-COCO 2014 (COCO14) (Lin et al., 2014), NUS-WIDE (Chua et al., 2009), and CUB200 (Wah et al., 2011). We reserve 20% of each original training set for validation and use ground-truth labels for both validation and testing. Following the complementary-label setting of Gao et al. (2026b), we generate complementary labels by uniformly sampling negative image–class pairs from the optimization training split. The sampling ratio $\rho \in \{0.1, 0.2, 0.3, 0.4\}$ denotes the fraction of all negative image–class pairs in this split selected as complementary labels. Dataset details are provided in Appendix C.

We compare CoDi with two MLCLL baselines, GDF (Gao et al., 2023) and CTL (Gao et al., 2025), and three VLM-based baselines: AEVLP (Tran et al., 2025), VLPL (Xing et al., 2024), and GR Loss with the DAMP pseudo-label adapter (Chen et al., 2024) (GR+DAMP). All methods use an ImageNet-pretrained ResNet-50 classifier, while the VLM-based baselines and CoDi additionally use information from pretrained CLIP.

CoDi uses a frozen CLIP ViT-B/16 teacher and AdamW optimization (Loshchilov & Hutter, 2019). At inference, only a single global forward pass of the EMA student is required. We set the per-GPU batch sizes to 64, 96, 96, and 64 and the maximum training epochs to 50, 30, 30, and 40 for VOC07, COCO14, NUS-WIDE, and CUB, respectively. Weight decay is 10^{-4}, and the learning rates are 10^{-5} for the backbone and 10^{-4} for the classifier and projection layers. We fix $w_{\text{wide}} = 0.5$ throughout the main comparison. The evidence weights $(\alpha_g, \alpha_r, \alpha_a, \alpha_e)$ are $(0.55, 0.45, 0, 0)$ during warm-up and $(0.35, 0.25, 0.30, 0.10)$ during joint training. We set $\xi = 0.05$, $\gamma_- = 2$, $\eta = 0.2$, $\beta = 0.6$, and $q = 0.7$ across all datasets. We set ζ to 0.9 for VOC07, 0.85 for CUB200, and 0.92 for the other two datasets. The maximum loss weights $(\lambda_p, \lambda_d, \lambda_c, \lambda_s, \lambda_n)$ are $(1.0, 0.5, 0.5, 0.05, 0.2)$, with $\lambda_d = 0$ during final fine-tuning. For each dataset, we repeat training with five random seeds and report the means and sample standard deviations (std) across the five runs. Here, we adopt five MLL criteria, including class-wise mean average precision (mAP), average precision, ranking loss,

Table 1: The results of mAP (mean±std, in %) on four datasets with $\rho \in \{0.1, 0.2, 0.3, 0.4\}$. The best result is shown in boldface, where •/◦ indicates whether CoDi is superior/inferior to baselines with pairwise t-test (at the 0.05 significance level).

Dataset	ρ	CTL	GDF	AEVLP	GR+DAMP	VLPL	CoDi
VOC07	0.1	29.68 ± 0.47•	29.99 ± 0.37•	83.06 ± 0.20	$\mathbf{83.62 \pm 0.32}$	75.60 ± 1.50•	82.17 ± 0.51
	0.2	44.58 ± 0.59•	43.74 ± 0.55•	82.74 ± 0.20•	$\mathbf{83.97 \pm 0.12}$	74.19 ± 1.18•	83.74 ± 0.47
	0.3	59.37 ± 1.28•	54.79 ± 0.77•	82.76 ± 0.19•	84.21 ± 0.11•	73.54 ± 0.95•	$\mathbf{85.57 \pm 0.14}$
	0.4	70.00 ± 1.31•	63.18 ± 0.69•	82.61 ± 0.90•	84.34 ± 0.15•	73.19 ± 1.28•	$\mathbf{86.53 \pm 0.14}$
COCO14	0.1	30.24 ± 0.54•	29.41 ± 0.27•	63.19 ± 0.27	$\mathbf{63.38 \pm 0.21}$	61.19 ± 0.19•	63.24 ± 0.03
	0.2	40.40 ± 0.60•	42.12 ± 0.43•	63.84 ± 0.33•	63.89 ± 0.17•	64.66 ± 0.24•	$\mathbf{65.57} \pm 0.05$
	0.3	50.36 ± 0.52•	51.35 ± 0.34•	64.30 ± 0.32•	64.21 ± 0.33•	65.73 ± 0.24•	$\mathbf{67.01 \pm 0.04}$
	0.4	58.73 ± 0.56•	58.11 ± 0.28•	64.66 ± 0.30•	64.50 ± 0.27•	65.12 ± 0.12•	$\mathbf{68.19} \pm 0.04$
NUS-WIDE	0.1	15.45 ± 0.17•	15.14 ± 0.29•	38.10 ± 0.14•	38.64 ± 0.09•	44.28 ± 0.24•	$\mathbf{46.65 \pm 0.09}$
	0.2	20.76 ± 0.70•	21.65 ± 0.06•	39.73 ± 0.35•	40.06 ± 0.18•	45.99 ± 0.31•	$\mathbf{48.33 \pm 0.02}$
	0.3	26.16 ± 0.75•	27.49 ± 0.22•	41.48 ± 0.27•	41.39 ± 0.27•	46.29 ± 0.41•	$\mathbf{49.16 \pm 0.04}$
	0.4	30.94 ± 0.60•	33.46 ± 0.24•	42.50 ± 0.37•	42.56 ± 0.28•	46.40 ± 0.29•	$\mathbf{50.03 \pm 0.03}$
CUB200	0.1	14.26 ± 0.10•	14.24 ± 0.10•	18.73 ± 0.19•	19.17 ± 0.33•	13.56 ± 0.26•	$\mathbf{21.82 \pm 0.07}$
	0.2	16.40 ± 0.06•	16.39 ± 0.08•	18.71 ± 0.04•	19.46 ± 0.17•	12.84 ± 0.08•	$\mathbf{22.96 \pm 0.11}$
	0.3	18.16 ± 0.05•	18.10 ± 0.06•	18.80 ± 0.23•	19.66 ± 0.20•	12.69 ± 0.15•	$\mathbf{23.62 \pm 0.08}$
	0.4	20.13 ± 0.04•	20.01 ± 0.03•	18.94 ± 0.22•	19.85 ± 0.18•	12.47 ± 0.17•	$\mathbf{24.23 \pm 0.10}$

one error, and normalized coverage, to measure the performance of approaches. For mAP and average precision, greater values indicate better performance, while for the remaining three criteria, smaller values indicate better performance. We implement our experiments in PyTorch (Paszke et al., 2019) and run them on two NVIDIA RTX 5090 GPUs. The code will be released upon acceptance.

4.2 Results

Table 1 reports the mAP results on four datasets under different complementary annotation rates, where the remaining results of four criteria are stated in Appendix C. CoDi achieves the highest mAP in 13 of the 16 settings, with statistically significant improvements over all baselines in these settings according to the reported pairwise t-tests. Specifically, CoDi outperforms CTL and GDF across all settings, which suggests that augmenting complementary supervision with pretrained positive semantic knowledge improves MLCLL performance. In addition, CoDi outperforms the strongest baseline in every setting on NUS-WIDE and CUB200, with gains of 2.34%–3.63% and 2.65%–4.10%, respectively. This advantages over the VLM-based baselines further suggest the effectiveness of integrating such knowledge with reliable negative constraints. Moreover, CoDi's performance improves as ρ increases from 0.1 to 0.4, consistent with its design: additional complementary labels provide more negative constraints to guide pseudo-positive learning, while CoDi retains positive evidence from the CLIP teacher. At $\rho = 0.4$, CoDi outperforms the strongest baseline on each dataset by 2.19%, 3.07%, 3.63%, and 4.10%, respectively. Nevertheless, GR+DAMP obtains higher mAP on VOC07 at $\rho \in \{0.1, 0.2\}$ and COCO14 at $\rho = 0.1$, which indicate that the advantage of CoDi depends on the dataset and the availability of complementary supervision.

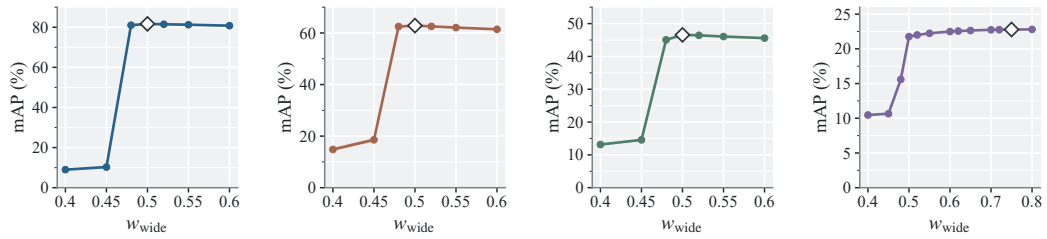
To verify the effectiveness of different strategies, we conduct ablation studies to evaluate the contributions of three key strategies in CoDi: class-aware views, median calibration, and soft-evidence updates. Specifically, we compare CoDi with three variants at $\rho = 0.1$: (1) CoDi w/o class-aware views, which removes the CAM-guided views from multi-view positive evidence construction and redistributes their fusion weight among the remaining branches; (2) CoDi w/o median calibration, which replaces the median-based wide calibration in equation 3 with centered calibration. Here, centered calibration obtains $q_{ic}^{(v)}$ by dividing $s_{ic}^{(v)}$ by the sum of $s_{ic}^{(v)}$ and the clipped class-wise global median used in equation 3, which retains the same median estimates and maps a score equal to the clipped median to 0.5; and (3) CoDi w/o soft-evidence updates, which uses a target of one instead of e_{ic}

Table 2: Ablation studies on mAP (mean±std, in %) at $\rho = 0.1$ over 5 trials. The best performance is shown in boldface, where $w_{\text{wide}} = 0.5$.

Variants	VOC07	COCO14	NUS-WIDE	CUB200
CoDi	$\mathbf{82.17 \pm 0.51}$	$\mathbf{63.24 \pm 0.03}$	$\mathbf{46.65 \pm 0.09}$	21.82 ± 0.07
CoDi w/o class-aware views	81.44 ± 0.44	62.77 ± 0.15	46.43 ± 0.05	20.52 ± 0.08
CoDi w/o median calibration	80.84 ± 0.57	62.28 ± 0.05	44.99 ± 0.32	18.49 ± 0.05
CoDi w/o soft-evidence updates	73.26 ± 1.33	58.25 ± 0.11	42.26 ± 0.08	$\mathbf{21.94 \pm 0.05}$

in the EMA bank update for each accepted pseudo-positive, while retaining EMA smoothing and the reliability-weight update. All other experimental settings, including models and hyperparameters, are kept identical to those of CoDi across all datasets, as described in subsection 4.1. From Table 2, the results show that class-aware views and median calibration improve mAP across all four datasets, with the largest gains on CUB200 (1.24% and 3.27% , respectively). Replacing soft-evidence with targets of one reduces mAP by 8.91%, 4.99%, and 4.39% on VOC07, COCO14, and NUS-WIDE. These results support the effectiveness of three strategies in CoDi, although their individual benefits vary across datasets.

We then explore the effect of varying w_{wide} values in equation 3 at $\rho = 0.1$. We evaluate $w_{\text{wide}} \in \{0.40, 0.45, 0.48, 0.50, 0.52, 0.55, 0.60\}$ on all datasets and additionally test $\{0.62, 0.65, 0.70, 0.72, 0.75, 0.80\}$ on CUB200. As shown in Figure 2, VOC07, COCO14, and NUS-WIDE exhibit their largest sampled improvement between 0.45 and 0.48, followed by relatively stable performance. CUB200 improves more gradually beyond 0.50, suggesting dataset-dependent calibration requirements. These results suggest that overly small scales may suppress useful positive evidence, while suitable scales provide a relatively stable operating range. Validation selects 0.50 for the first three datasets and 0.75 for CUB200. Here, we use $w_{\text{wide}} = 0.50$ across all datasets in the main comparison to maintain a consistent evaluation protocol. Due to the page limitation, the analysis of the remaining hyperparameters is shown in Appendix C.



(a) VOC07 (b) COCO14 (c) NUS-WIDE (d) CUB200

Figure 2: Effect of w_{wide} on performance across four datasets at $\rho = 0.1$. Hollow diamonds indicate the values selected using the validation sets.

5 Conclusion

In this paper, we proposed CoDi, a CLIP-based framework for MLCLL. Our analysis showed that non-exhaustive complementary labels can leave residual label uncertainty and that negative-only learning can admit degenerate solutions. To address these limitations, CoDi combines multi-view evidence from a frozen CLIP teacher with reliability-aware pseudo-positive learning, which supplements observed negative constraints with positive supervision. The resulting knowledge is distilled into a standalone student, enabling CLIP-free inference. Experiments across multiple benchmarks demonstrate the effectiveness of CoDi and highlight the value of integrating pretrained semantic knowledge with complementary supervision for MLCLL.

References

Rabab Abdelfattah, Qing Guo, Xiaoguang Li, Xiaofeng Wang, and Song Wang. Cdul: Clip-driven unsupervised learning for multi-label image classification. In Proceedings of the 19th IEEE/CVF International Conference on Computer Vision, pp. 1348–1357, Paris, France, 2023.

Cheng Chen, Yifan Zhao, and Jia Li. Adapting dense vision-language relationships for multi-label classification with partial label. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2026a.

Cheng Chen, Jingyu Zhou, Yifan Zhao, and Jia Li. Adapting vision-language models from iconic to inclusive for multi-label recognition without labels. International Journal of Computer Vision, 134(6):294, 2026b.

Yanxi Chen, Chunxiao Li, Xinyang Dai, Jinhuan Li, Weiyu Sun, Yiming Wang, Renyuan Zhang, Tinghe Zhang, and Bo Wang. Boosting single positive multi-label classification with generalized robust loss. In Proceedings of the 33rd International Joint Conference on Artificial Intelligence, pp. 3825–3833, Jeju, Korea, 2024.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: A real-world web image database from National University of Singapore. In Proceedings of the 8th ACM International Conference on Image and Video Retrieval, pp. 1–9, Santorini, Greece, 2009.

Zixuan Ding, Ao Wang, Hui Chen, Qiang Zhang, Pengzhang Liu, Yongjun Bao, Weipeng Yan, and Jungong Han. Exploring structured semantic prior for multi label recognition with incomplete labels. In Proceedings of the 34th IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3398–3407, Vancouver, Canada, 2023.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. International Journal of Computer Vision, 88(2):303–338, 2010.

Yi Gao, Miao Xu, and Min-Ling Zhang. Unbiased risk estimator to multi-labeled complementary label learning. In Proceedings of the 32nd International Joint Conference on Artificial Intelligence, pp. 3732–3740, Macao, China, 2023.

Yi Gao, Miao Xu, and Min-Ling Zhang. Complementary to multiple labels: A correlation-aware correction approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12):9179–9191, 2024.

Yi Gao, Jing-Yi Zhu, Miao Xu, and Min-Ling Zhang. Multi-label learning with multiple complementary labels. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(9):8013–8024, 2025.

Yi Gao, Davide Anguita, Roli Fabio, Min-Ling Zhang, Fulvio Mastrogiovanni, and Luca Oneto. Multi-label complementary labels learning with hard logical constraints. In Proceedings of the 34th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, pp. 81–86, Bruges, Belgium, 2026a.

Yi Gao, Yuan-Yuan Meng, Miao Xu, and Min-Ling Zhang. Data augmentation with regularization for multi-labeled complementary label learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 48(3):2368–2379, 2026b.

Bozhen Hu, Cheng Tan, Yongjie Xu, Zhangyang Gao, Jun Xia, Lirong Wu, and Stan Z Li. Protgo: Function-guided protein modeling for unified representation learning. In Advances in Neural Information Processing Systems 37, pp. 88581–88604, Vancouver, Canada, 2024.

Ping Hu, Ximeng Sun, Stan Sclaroff, and Kate Saenko. Dualcoop++: Fast and effective adaptation to multi-label recognition with limited annotations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(5):3450–3462, 2023.

Dongseob Kim and Hyunjung Shim. Classifier-guided CLIP distillation for unsupervised multi-label classification. In Proceedings of the 38th IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4661–4671, Nashville, TN, 2025.

Maxat Kulmanov, Francisco J Guzmán-Vega, Paula Duek Roggli, Lydie Lane, Stefan T Arold, and Robert Hoehndorf. Protein function prediction as approximate semantic entailment. Nature Machine Intelligence, 6(2):220–228, 2024.

Kun Li, Pengyu Liu, Dan Guo, Fei Wang, Zhiliang Wu, Hehe Fan, and Meng Wang. Mmad: Multi-label micro-action detection in videos. In Proceedings of the 20th IEEE/CVF International Conference on Computer Vision, pp. 13225–13236, Honolulu, HI, 2025.

Zhifang Liao, Junhao Li, HaoKang Ding, and Yucheng Song. Darc: Dual adjustment reasoning with counterfactuals for trustworthy chest x-ray classification. In Proceedings of the 39th IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 28234–28243, Denver, CO, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In Proceedings of the 13th European Conference on Computer Vision, pp. 740–755, Zurich, Switzerland, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2019.

Gengyu Lyu, Bohang Sun, Xiang Deng, and Songhe Feng. Addressing multi-label learning with partial labels: from sample selection to label selection. In Proceedings of the 39th AAAI Conference on Artificial Intelligence, volume 39, pp. 19251–19259, Philadelphia, PA, 2025.

Dwarikanath Mahapatra, Antonio Jimeno Yepes, Behzad Bozorgtabar, Sudipta Roy, Zongyuan Ge, and Mauricio Reyes. Multi-label generalized zero shot chest x-ray classification by combining image-text information with feature disentanglement. IEEE Transactions on Medical Imaging, 44(1):31–43, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Ze-Ming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Jun-Jie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems 32, pp. 8024–8035, Vancouver, Canada, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, pp. 8748–8763, Virtual Event, 2021.

Samyak Rawlekar, Shubhang Bhatnagar, and Narendra Ahuja. PositiveCoOp: Rethinking prompting strategies for multi-label recognition with partial annotations. In Proceedings of the 25th IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5863–5872, Tucson, AZ, 2025.

Dengdi Sun, Hongxing Xie, Zhendong Cai, Leilei Ma, and Bin Luo. Visual–label alignment and attribute-aware prompt for multi-label image recognition with partial labels. ACM Transactions on Multimedia Computing, Communications and Applications, 22(8):1–24, 2026.

Ximeng Sun, Ping Hu, and Kate Saenko. DualCoOp: Fast adaptation to multi-label recognition with limited annotations. In Advances in Neural Information Processing Systems 35, pp. 30569–30582, New Orleans, LA, 2022.

Yi Tang, Yi Gao, Yonggang Luo, Jucheng Yang, Miao Xu, and Min-Ling Zhang. Unlearning from weakly supervised learning. In Proceedings of the 33rd International Joint Conference on Artificial Intelligence, pp. 5000–5008, Jeju, South Korea, 2024.

Luong Tran, Thieu Vo, Anh Nguyen, Sang Dinh, and Van Nguyen. More reliable pseudo-labels, better performance: A generalized approach to single positive multi-label learning. In Proceedings of the 20th IEEE/CVF International Conference on Computer Vision, pp. 1349–1358, Honolulu, HI, 2025.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Wenkang Wang, Yunyan Shuai, Min Zeng, Wei Fan, and Min Li. Dpfunc: accurately predicting protein function via deep learning with domain-guided structure information. Nature communications, 16(1):70, 2025.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the 30th IEEE/CVF conference on computer vision and pattern recognition, pp. 2097–2106, Honolulu, HI, 2017.

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 37th IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7799–7808, Seattle, WA, 2024.

Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering, 26(8):1819–1837, 2014.

Jing-Yi Zhu, Yi Gao, Miao Xu, and Min-Ling Zhang. Comrank: Ranking loss for multi-label complementary label learning. In Advances in Neural Information Processing Systems 38, pp. 120011–120037, San Diego, CA, 2025.

Yuhan Zhu, Guozhen Zhang, Jing Tan, Gangshan Wu, and Limin Wang. Dual detrs for multi-label temporal action detection. In Proceedings of the 37th IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18559–18569, Seattle, WA, 2024.