RESEARCH ARTICLE

Reliability-Aware Active Crowdsourcing and Ontology-Constrained Multi-Evidence Clustering for Agricultural Multimedia Knowledge-Base Construction

Hongda Li[1], **Zixin Zhu**[1], **Guanxu Ren**[1], **Yisheng Miao**[1,2,3,4] **and Peng Hao**[1,2,3,4,*]

[1] National Engineering Research Center for Information Technology in Agriculture, Beijing 100101, China
[2] Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China
[3] Key Laboratory of Digital Village Technology, Ministry of Agriculture and Rural Affairs, Beijing 100125, China
[4] Beijing Research Institute of Agricultural Artificial Intelligence and Robotics, Beijing 100097, China

Abstract

Constructing agricultural multimedia knowledge bases is complicated by heterogeneous evidence, noisy crowdsourced annotations, variations in sample difficulty, and inadequate integration of multisource information. Using tomato disease and pest data, this study proposes a unified framework comprising Multi-evidence Reliability-Aware Active Crowdsourcing (MRAC) and Ontology-Constrained Multi-Evidence Clustering and Refinement (OMCR). MRAC jointly models annotator reliability, sample difficulty, content-evidence quality, and evidence availability. It selects sample-annotator pairs according to the expected information gain per unit cost and produces calibrated soft-label posteriors. OMCR integrates frozen visual features, crowdsourcing posteriors, and agricultural ontology constraints, and uses class prototypes, pairwise constraints, and risk scores for sample clustering, mislabeled-instance detection, and label refinement. Under 30% synthetic annotation noise, MRAC achieved an Accuracy of 0.9732 and a Macro-F1 of 0.9676. MRAC-EIG achieved the highest mean Macro-F1 (0.9631) and the lowest mean relative annotation cost (1.0080) among the evaluated policies. The complete OMCR model achieved ACC, NMI, and ARI values of 0.9708, 0.9293, and 0.9500, respectively. The unified framework achieved Precision@5, Precision@10, and nDCG@10 values of 0.9251, 0.9223, and 0.9233, respectively, with a 14.97% relative improvement in nDCG@10 over visual-only retrieval. These results show that the proposed framework improves noisy-label aggregation, multi-evidence knowledge organization, and retrieval ranking while controlling annotation costs, offering a practical route to reliable agricultural multimedia knowledge-base construction.

Keywords: agricultural multimedia knowledge construction, tomato diseases and pests, active crowdsourcing, multimodal reliability awareness, ontology-constrained clustering, image retrieval.

Submitted: 09 May 2026
Revised: 03 November 2026
Accepted: 05 November 2026
Published: 10 November 2026

Vol. 2, **No.** 4, 2026.
10.62762/DIA.2026.512829

***Corresponding author:**
✉ Peng Hao
haop007@nercita.org.cn

Citation
Li, H., Zhu, Z., Ren, G., Miao, Y., & Hao, P. (2026). Reliability-Aware Active Crowdsourcing and Ontology-Constrained Multi-Evidence Clustering for Agricultural Multimedia Knowledge-Base Construction. *Digital Intelligence in Agriculture*, 2(4), 214–229.

© 2026 by the Authors. Published by Institute of Central Computation and Knowledge. This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

1 Introduction

The increasing adoption of intelligent sensing and agricultural information technologies has generated large volumes of multimedia data from diverse sources throughout agricultural production [1]. These data contain valuable knowledge concerning crop growth and the occurrence and progression of diseases and pests [2]. Effective knowledge-oriented organization of such data is fundamental to improving intelligent agricultural diagnosis and decision support [3, 4]. Existing studies have primarily established task-specific mappings between visual features and predefined categories [5], with limited modeling of cross-modal semantic associations. As multimodal intelligence becomes increasingly integrated with knowledge engineering, agricultural intelligence is expanding from visual recognition alone to visual question answering, diagnosis with multimodal large models, and knowledge services driven by knowledge graphs [6]. Converting distributed, modality-heterogeneous, and variable-quality agricultural multimedia data into semantically consistent, explicitly related, and trustworthy structured knowledge has therefore become a critical challenge at the intersection of agricultural knowledge engineering and multimodal intelligence.

The complexity of agricultural multimedia data arises not only from their structure but also from differences in how agricultural knowledge is expressed across media types. The objective of agricultural multimedia analysis is shifting from visual category recognition toward content understanding and the construction of semantic relationships among symptoms, crops, pathogens, and disease concepts. To address the limited ability of image classification to represent specialized knowledge, Agri-LLaVA incorporates knowledge of agricultural diseases and pests to improve visual question answering and domain-specific explanations [7]. CDDM combines crop-disease images with diagnostic question-answer pairs, extending the task from category recognition to symptom description and diagnostic reasoning [8]. AgriDoctor further adopts a modular framework that coordinates classification, detection, retrieval, and large models [9], using vision-language models and domain knowledge to improve agricultural content understanding. The crop-specific TomaMMU benchmark likewise shows that agricultural datasets are evolving from single-label annotations toward visual question answering and domain-knowledge comprehension [10]. Nevertheless, existing studies largely depend on static image-text data and curated high-quality labels, with insufficient attention to temporal information in videos or to the conversion of raw multimedia data into trustworthy knowledge. Reliable annotation at a reasonable cost is therefore the first challenge in agricultural multimedia knowledge construction.

Agricultural multimedia knowledge construction depends on manual annotation. However, domain experts are scarce and costly, whereas nonexpert annotators may produce noisy labels because of differences in expertise. Crowdsourcing can increase the scale of data production, but annotation efficiency and result reliability must be considered jointly. Zhang et al. [11] proposed an active crowdsourcing method that iteratively queries noisy labels from crowd workers to reduce redundant annotation and improve multilabel classification. Crowd-Certain estimates annotator reliability from response consistency and aggregates multiple labels using reliability-based weights [12]. ActiveLab actively chooses between annotating new samples and relabeling existing ones, directing a limited budget toward highly uncertain data [13]. These studies improve crowdsourcing by reducing per-task cost, limiting the influence of low-quality labels, or optimizing budget allocation. However, existing methods commonly represent annotator competence with a global weight and treat sample selection, task assignment, and label aggregation as separate processes. This design is poorly suited to the domain specificity of agricultural categories and the heterogeneous difficulty of multimedia samples.

Multimedia content, annotation results, and agricultural ontologies must subsequently be integrated to support unified knowledge organization, denoising, and refinement. Multiview clustering research has begun to distinguish information shared across modalities from modality-specific information. MFLVC establishes cross-view consistency at multiple levels, including low-level features, high-level semantics, and cluster assignments, thereby reducing interference caused by the direct fusion of view-specific noise [14]. ICMVC uses high-confidence guidance and attention-based fusion to address missing views and incomplete representations [15]. Multimedia aggregation has therefore progressed beyond feature concatenation toward learning unified cross-modal semantic structures through contrastive learning, attention weighting, and confidence guidance. In

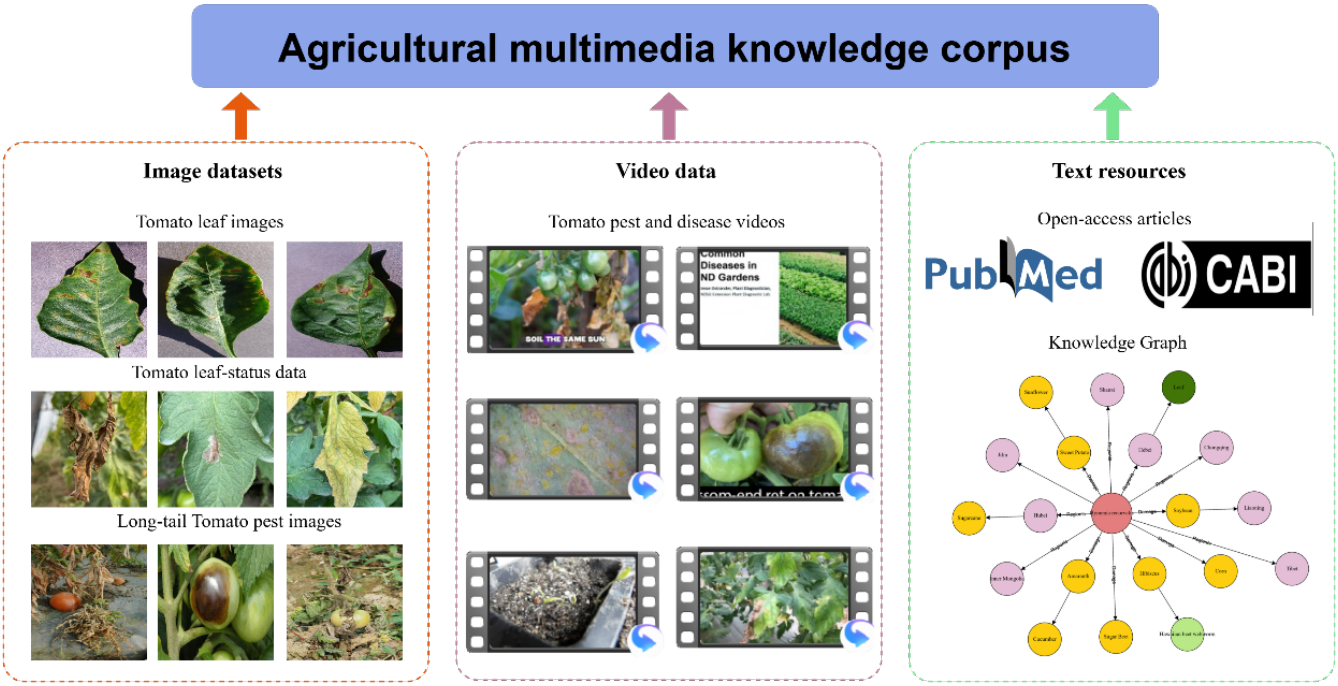


Figure 1. Composition and data sources of the agricultural multimedia knowledge corpus.

parallel, in the agricultural domain, ontology-guided knowledge graph approaches integrate named entity recognition and structured relation extraction to organize domain-specific knowledge, as exemplified by the tomato leaf pest and disease knowledge graph [16], whereas CropDP-KG provides a domain-knowledge foundation for organizing concepts, attributes, and relations associated with crop diseases and pests in China [17]. However, multimedia clustering and knowledge-graph research remain largely disconnected and cannot simultaneously refine known class structures and correct noisy labels in raw multimedia data. Crowdsourcing confidence and agricultural ontology constraints have not yet been incorporated into a unified clustering process, limiting the semantic consistency and reliability of knowledge aggregation.

To address these limitations, this study develops a unified workflow for processing agricultural multimedia knowledge. First, a reliability-aware active crowdsourcing method jointly selects informative samples and suitable annotators to obtain aggregated labels with probabilistic confidence. Second, an ontology-guided multi-evidence clustering method integrates visual content features, crowdsourced soft-label posteriors, and agricultural ontology constraints, and uses agricultural concept hierarchies and relational constraints to identify and correct noisy labels. Finally, the proposed methods are evaluated on multimedia data for tomato diseases and pests. The principal contributions are as follows:

(1) A reliability-aware active crowdsourcing method is proposed. It jointly considers sample information value and annotator reliability, actively matches informative samples with suitable annotators, and produces confidence-aware aggregated labels.

(2) An ontology-constrained multi-evidence clustering and refinement method is developed. Content features, crowdsourced soft-label posteriors, and agricultural ontology constraints are represented as complementary evidence views. Class prototypes and ontology-derived pairwise constraints are then used for sample clustering, noise detection, and label refinement.

(3) A unified workflow is established for agricultural multimedia knowledge construction. Using tomato disease and pest data, the workflow connects crowdsourced annotation, clustering refinement, and knowledge organization, providing methodological support for converting heterogeneous agricultural data into structured and trustworthy knowledge.

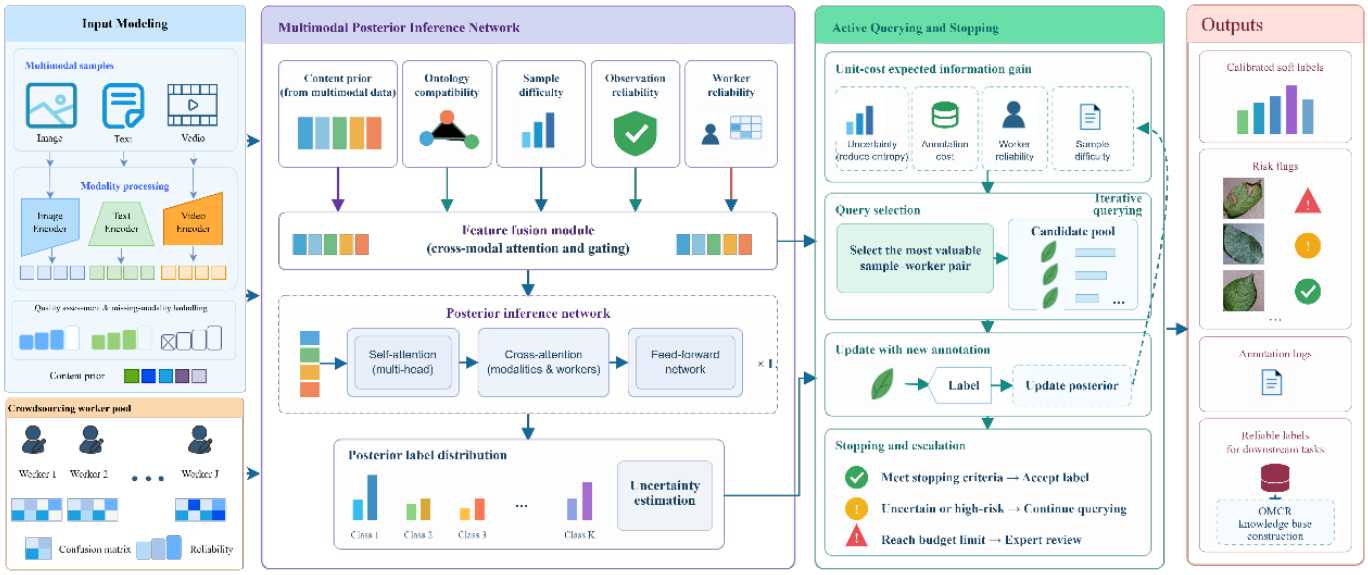


Figure 2. Architecture of the proposed MRAC framework for reliability-aware active annotation of agricultural multimedia data.

2 Materials and Methods

2.1 Data collection

This study considers multimedia knowledge construction for tomato diseases and pests. The data sources include tomato leaf images from PlantVillage [18], tomato leaf condition data, long-tailed images of tomato pests, videos, open-access articles from PubMed Central (PMC), candidate terms from AGROVOC, the Plant Ontology [19], the Trait Ontology, and CropDP-KG [17]. These data were used for visual representation learning, open-set evaluation, exploratory video analysis, textual knowledge extraction, and ontology-constraint construction.

As illustrated in Figure 1, the dataset comprised 42,775 files with a total size of approximately 1.47 GB. The PlantVillage subset contained 16,011 tomato leaf images, including 12,803 images in the original training set and 3,208 in the test set. After a validation set was separated from the original training data, the final training, validation, and test sets used for model development contained 10,882, 1,921, and 3,208 images, respectively. No duplicate images were found across the three sets.

2.2 Frozen visual representations and baselines

To ensure that MRAC and OMCR use a consistent source of underlying visual evidence, a frozen ResNet-18 encoder [20] pretrained on ImageNet was used to extract 512-dimensional visual features from the global average-pooling layer, followed by L2 normalization. A multinomial logistic-regression classification head was then fitted to the PlantVillage training set only, and its regularization parameter C_{reg} was selected from $\{0.1, 1, 10\}$ on the validation set; the test set was not used for parameter selection. The resulting class probabilities and visual features were used in the subsequent MRAC and OMCR experiments.

In addition to this visual-prior model, a ResNet-18 model with a fine-tuned layer4 block was trained for an independent analysis of visual classification performance. Training used only the PlantVillage training set, and the best checkpoint was selected according to the validation-set Macro-F1.

To calibrate the class probabilities, temperature scaling [21] was applied to the validation-set logits. The temperature-scaled class probabilities and the optimal temperature parameter T^* were defined as follows:

$$T^* = \arg\min_{T>0} \; -\frac{1}{N_{\mathrm{v}}} \sum_{i=1}^{N_{\mathrm{v}}} \log p_T(y_i \mid x_i), \qquad (1)$$
$$p_T(y \mid x) = \mathrm{softmax}\big(f(x)/T\big),$$

where N_{v} denotes the number of validation samples, $f(x)$ is the uncalibrated logit vector, and $p_T(y \mid x)$ denotes the temperature-scaled class probabilities. After model selection and probability calibration, the classification-head parameters and temperature parameter were fixed. The calibrated logistic-regression visual prior based on frozen

features subsequently provided shared content features and class priors for MRAC and OMCR.

2.3 Multi-evidence reliability-aware active crowdsourcing

This study proposes the Multi-evidence Reliability-Aware Active Crowdsourcing method illustrated in Figure 2. Under a constrained annotation budget, MRAC jointly models the sample-specific content prior, annotator reliability, sample difficulty, modality quality, and annotation cost, and produces calibrated soft-label posteriors. Unlike conventional majority voting or aggregation based solely on annotator competence, MRAC incorporates multimodal evidence quality and differences among annotators into posterior inference. During task selection, MRAC selects informative sample-annotator pairs according to the expected information gain per unit cost, improving label aggregation while controlling annotation expenditure. Samples that do not satisfy the predefined confidence criteria within the available budget are designated high-risk instances and referred for expert review.

2.3.1 *Hierarchical posterior model*

For the ith sample, let x_i^m denote the observed content of modality $m \in \{\text{img}, \text{txt}, \text{vid}\}$, $M_i^m \in \{0, 1\}$ indicate whether that modality is present, and s_i^m denote the quality measure for modality m. Let y_i represent the latent true class, z_{ij} denote the label assigned by annotator j to sample i, and Z_i denote all crowdsourced labels collected for sample i. The MRAC posterior is defined as:

$$p(y_i \mid x_i, Z_i, G) \propto p_\phi(y_i \mid x_i)\, \psi_G(y_i) \times \prod_{j \in A_i} p(z_{ij} \mid y_i, \Theta_j, d_i, r_i), \quad (2)$$

where G denotes the agricultural ontology; $p_\phi(y_i \mid x_i)$ is the content prior, with parameters ϕ, obtained by combining the frozen visual prior with available textual and video evidence; $\psi_G(y_i)$ is the soft compatibility factor supplied by the agricultural ontology; Θ_j is the confusion matrix of annotator j; d_i denotes sample difficulty; r_i represents the observational reliability of the sample; and A_i is the set of annotators who have provided valid labels for sample i. This formulation retains class-specific annotation biases instead of reducing annotator reliability to a single competence score. To derive the ontology compatibility factor, each candidate class label is mapped to an agricultural ontology concept. Preferred and alternative terms in AGROVOC are used to resolve naming variants, while relations to the relevant crop, disease, and pest concepts determine a class-level compatibility weight.

2.3.2 *Evidence availability and quality normalization*

The missing-modality status was recorded for each sample. Overall multimodal quality Q_i was calculated using missingness-aware weighted normalization:

$$Q_i = \frac{\sum_m M_i^m w_m s_i^m}{\sum_m M_i^m w_m + \varepsilon}, \quad (3)$$

where w_m is the weight assigned to modality m, and ε is a numerical-stability term that prevents division by zero. Only modalities that are actually present contribute to quality aggregation. A missing modality is not replaced with a zero-quality value, which prevents modality absence from being misinterpreted as poor quality.

Cross-modal inconsistency was calculated only when at least two modalities were available. For image-only samples, cross-modal inconsistency was treated as unobserved rather than using missing modalities to classify the sample as either high or low risk. The number of available modalities $n_i^{\text{obs}} = \sum_m M_i^m$ was included explicitly in the sample-difficulty estimator and active-query model.

2.3.3 *Active-query criterion and stopping mechanism*

After estimating the initial class posterior, MRAC uses an expected-information-gain strategy, denoted MRAC-EIG, to prioritize crowdsourced labels that are expected to reduce mean uncertainty in class decisions under a limited annotation budget. Unlike methods that select samples solely according to model uncertainty, MRAC-EIG jointly accounts for current posterior uncertainty, the response reliability of candidate annotators, and the cost of each annotation. The next query is therefore determined at the level of the sample-annotator pair.

Given the current class posterior $q_i(y)$, annotator reliability, and annotation cost c_j, MRAC calculates the expected information gain (EIG) per unit cost for every eligible sample-annotator pair. The optimal query pair is defined as:

$$(i^*, j^*) = \arg\max_{(i,j) \in \mathcal{C}} \frac{H[q_i] - \mathbb{E}_{z \sim p(z \mid i,j)} H[q_i(y \mid z)]}{c_j + \varepsilon}, \quad (4)$$

where $H[\cdot]$ denotes information entropy, $\mathcal{C}$ is the set of eligible candidate query pairs, z is the potential response of annotator j to sample i, and $p(z \mid i, j)$ is the

corresponding predictive response distribution. The numerator represents the expected reduction in class uncertainty after one query, whereas the denominator normalizes query utility for differences in annotation cost. MRAC-EIG prioritizes pairs with high current posterior uncertainty that can be effectively resolved by reliable annotators and provide high information gain per unit cost.

At each iteration, the algorithm selects only the highest-scoring eligible pair, obtains its annotation, and updates the posterior. Further querying of a sample stops once the sample satisfies a stopping criterion jointly defined by posterior confidence, information entropy, and calibration risk. Samples that do not meet the criterion remain in the candidate pool. Those that still fail to meet it after the budget is exhausted are referred for expert review.

2.4 Ontology-constrained multi-evidence clustering and refinement

This study proposes Ontology-Constrained Multi-Evidence Clustering and Refinement (OMCR), whose overall architecture is shown in Figure 3. OMCR integrates heterogeneous multimedia content, crowdsourced annotation posteriors, and agricultural ontology knowledge to support sample clustering, label refinement, and risk-based routing within a unified framework. It treats content features extracted by the frozen encoder, soft-label posteriors produced by MRAC, and ontology-compatibility information as three evidence views, which are adaptively fused according to their availability and reliability.

OMCR uses reliable soft labels to construct class prototypes that guide the learning of cluster structure. It jointly constrains within-cluster compactness, cross-view consistency, agreement with crowdsourcing posteriors, and ontology compatibility to obtain stable sample representations and cluster assignments. After clustering, OMCR combines uncertainty in the crowdsourcing posterior, the distance between a sample and its class prototype, multiview disagreement, ontology conflicts, and local outlier scores into an integrated risk score. Thresholds determined on the validation set are then used for knowledge refinement and risk-based routing. Low-risk samples undergo automatic label alignment or near-duplicate candidate detection; medium-risk samples enter a queue for supplementary annotation or additional data collection; and high-risk samples are referred for expert review. Clusters that are internally compact but weakly similar to existing class

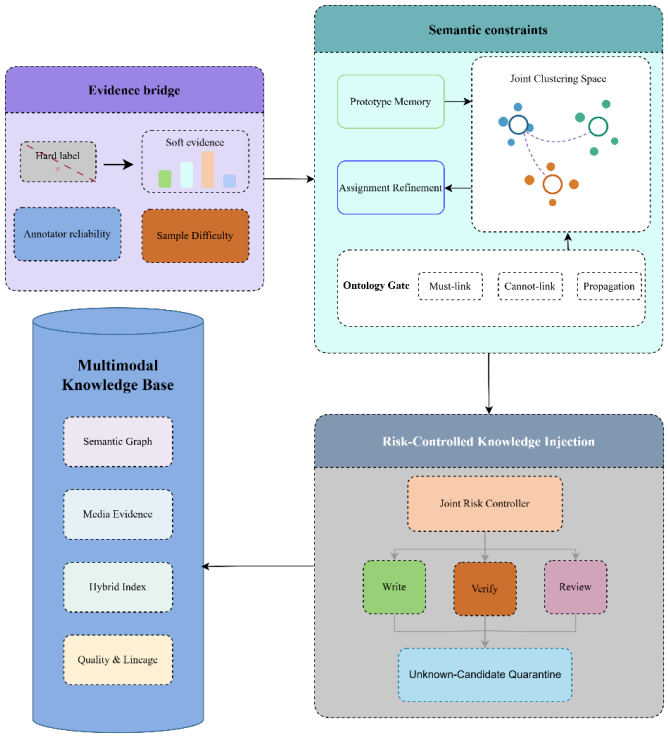


Figure 3. Overall framework of the proposed OMCR method for ontology-guided multi-evidence clustering and knowledge refinement.

prototypes are marked only as candidate unknown classes and are not directly assigned a new disease or pest category.

2.4.1 *Evidence views and adaptive fusion*

OMCR comprises three evidence views: the content view, crowdsourcing-posterior view, and ontology view, denoted by h_i^c, h_i^r, and h_i^o, respectively. Let $M_i^v \in \{0,1\}$ indicate whether view v is available, and let κ_i^v denote its quality or confidence weight. Let $P_v(\cdot)$ be the projection function that maps view v into a shared representation space. The fused representation is defined as:

$$z_i = \frac{\sum_{v\in\{c,r,o\}} M_i^v \kappa_i^v P_v(h_i^v)}{\sum_v M_i^v \kappa_i^v + \varepsilon}, \tag{5}$$

Only available evidence views participate in fusion; consequently, different samples may contain different evidence combinations. The principal quantitative experiments therefore follow a multi-evidence view-learning paradigm and do not assume that every sample contains strictly paired image, text, and video data.

2.4.2 Clustering objective and ontology constraints

The overall OMCR optimization objective is defined as:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{cluster}} + \lambda_v \mathcal{L}_{\text{align}} + \lambda_c \mathcal{L}_{\text{crowd}} \\ + \lambda_o \mathcal{L}_{\text{ontology}} + \lambda_r \mathcal{L}_{\text{robust}},\end{aligned} \tag{6}$$

where $\mathcal{L}_{\text{cluster}}$ promotes within-cluster compactness; $\mathcal{L}_{\text{align}}$ aligns evidence views that are simultaneously available for the same sample; $\mathcal{L}_{\text{crowd}}$ constrains the clustering distribution to agree with the soft-label posterior produced by MRAC; $\mathcal{L}_{\text{ontology}}$ introduces pairwise relations supported by the agricultural ontology; and $\mathcal{L}_{\text{robust}}$ reduces the influence of low-quality or highly noisy samples on the class prototypes. The coefficients λ_v, λ_c, λ_o, and λ_r control the relative contributions of the corresponding loss terms.

The ontology-based pairwise-constraint loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{ontology}} = \sum_{(a,b)\in ML} \|z_a - z_b\|_2^2 \\ + \sum_{(a,b)\in CL} \max\big(0, \mu - \|z_a - z_b\|_2\big)^2,\end{aligned} \tag{7}$$

Here, ML and CL denote the sets of must-link and cannot-link constraints, respectively, and μ is the minimum separation margin imposed on cannot-link pairs. Must-link and cannot-link constraints were generated exclusively from general agricultural ontologies, observed attributes, duplicate groups, and reliable annotations.

2.4.3 Integrated risk scoring and knowledge-refinement routing

For predicted class $\hat{c}_i$ and sample i, the integrated risk score is defined as:

$$\begin{aligned}R_i = \alpha H(q_i) + \beta\big[1 - \cos(h_i^c, \mathbf{p}_{\hat{c}_i})\big] \\ + \gamma D_i^{\text{mv}} + \delta O_i + \eta V_i,\end{aligned} \tag{8}$$

where $H(q_i)$ denotes the entropy of the crowdsourced soft-label posterior q_i; $\mathbf{p}_{\hat{c}_i}$ is the prototype of the predicted class, and $\cos(h_i^c, \mathbf{p}_{\hat{c}_i})$ is the cosine similarity between the sample content feature and this prototype; D_i^{mv} denotes multiview disagreement; O_i represents the degree of ontology conflict; and V_i is the local outlier score. The coefficients α, β, γ, δ, and η are the weights of the corresponding risk terms. A larger R_i indicates greater overall sample risk.

2.5 Evaluation metrics

2.5.1 Evaluation of crowdsourced label quality and annotation efficiency

During crowdsourced label acquisition, performance was evaluated in terms of aggregated-label correctness, class-balanced performance, probability calibration, and annotation-resource consumption. The metrics were label Accuracy, Macro-F1, Recall@3, expected calibration error (ECE), and mean annotation cost. Recall@3 is the proportion of samples whose true class is among the three classes with the highest posterior probability.

Let N denote the number of test samples, and let y_i and $\hat{y}_i$ denote the true and aggregated predicted classes of the ith sample, respectively. Label Accuracy is defined as:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i), \tag{9}$$

Let TP_c, FP_c, and FN_c denote the numbers of true positives, false positives, and false negatives for class c, respectively. The precision and recall for class c are $\text{Precision}_c = TP_c/(TP_c + FP_c)$ and $\text{Recall}_c = TP_c/(TP_c + FN_c)$, respectively, and Macro-F1 is defined as:

$$\begin{aligned}\text{Macro-F1} &= \frac{1}{C}\sum_{c=1}^{C} F1_c \\ &= \frac{1}{C}\sum_{c=1}^{C} \frac{2 \times \text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c},\end{aligned} \tag{10}$$

where C is the number of classes. Unlike overall Accuracy, Macro-F1 assigns equal weight to every class and is therefore less affected by class imbalance. A lower mean annotation cost indicates that fewer annotation resources are required to obtain reliable labels. This metric should be interpreted together with Accuracy or Macro-F1 to ensure that label quality is not reduced merely by issuing fewer queries.

The probabilistic reliability of the soft-label posterior was assessed using ECE, which measures the discrepancy between predictive confidence and empirical accuracy. The N samples were partitioned into B nonoverlapping bins according to predictive confidence, and the set of samples in the bth bin was denoted by $\mathcal{B}_b$. Thus,

$$\text{ECE} = \sum_{b=1}^{B} \frac{|\mathcal{B}_b|}{N}\big|\text{acc}(\mathcal{B}_b) - \text{conf}(\mathcal{B}_b)\big|, \tag{11}$$

where $\text{acc}(\mathcal{B}_b)$ is the empirical prediction accuracy within the bin, and $\text{conf}(\mathcal{B}_b)$ is the mean predictive

confidence of samples in that bin. An ECE closer to zero indicates better probability calibration of the soft-label posterior.

The experiments used reproducible, dimensionless relative costs. For each crowdsourcing random seed s, 32 costs were sampled independently: $c_j^{(s)} \sim \mathrm{Uniform}(0.75, 1.35)$. Under the same seed, all four noise-rate settings shared the same cost vector, which was independent of annotator competence. The annotation cost for an individual sample was defined as the sum of the costs of all queries issued for that sample and was averaged over all evaluation samples:

$$\mathrm{cost} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j\in\mathcal{J}_i} c_j, \tag{12}$$

where $\mathcal{J}_i$ denotes the set of annotators actually queried for sample i.

2.5.2 *Evaluation of clustering and knowledge-refinement quality*

Clustering performance and refinement quality were evaluated using clustering accuracy (ACC), normalized mutual information (NMI), adjusted Rand index (ARI), and the area under the precision-recall curve (AUPRC) for mislabeled-instance detection.

Because cluster identifiers produced by unsupervised clustering have no fixed correspondence with the ground-truth class labels, the predicted clusters were matched to the true classes through optimal label mapping before ACC was calculated:

$$\mathrm{ACC} = \max_{\pi}\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\big[y_i = \pi(u_i)\big], \tag{13}$$

where u_i denotes the predicted cluster identifier of sample i, and π is the optimal mapping from cluster identifiers to ground-truth class labels. The optimal mapping was obtained using the Hungarian algorithm. A higher ACC indicates stronger agreement between the clustering result and the true classes. NMI measures the information shared between the ground-truth partition Y and the predicted clustering partition U, and is defined as:

$$\mathrm{NMI}(Y,U) = \frac{2I(Y;U)}{H(Y)+H(U)}, \tag{14}$$

where $I(Y;U)$ is the mutual information between Y and U, and $H(Y)$ and $H(U)$ are their respective entropies. A higher NMI indicates greater agreement between the predicted clustering structure and the ground-truth class structure. ARI evaluates agreement between the ground-truth and predicted partitions from the perspective of sample pairs and corrects for agreement expected by chance. Let n_{ij} denote the number of samples shared by true class i and predicted cluster j, with $a_i = \sum_j n_{ij}$ and $b_j = \sum_i n_{ij}$. Then,

$$\mathrm{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - E}{\frac{1}{2}\left[\sum_i\binom{a_i}{2} + \sum_j\binom{b_j}{2}\right] - E}, \tag{15}$$

with $E = \sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\Big/\binom{N}{2}$. An ARI closer to 1 indicates stronger agreement between the predicted clustering and the ground-truth partition. The ability of the knowledge-refinement process to detect incorrect labels was evaluated using AUPRC, defined as:

$$\mathrm{AUPRC} = \int_0^1 \frac{TP(r)}{TP(r)+FP(r)}\,dr, \tag{16}$$

where r denotes recall, and $TP(r)$ and $FP(r)$ are the corresponding numbers of true positives and false positives, respectively. An AUPRC closer to 1 indicates better detection of incorrect labels across different risk thresholds.

2.5.3 *Evaluation of knowledge-service performance*

The application performance of the resulting agricultural multimedia knowledge base was evaluated for disease and pest diagnosis and for knowledge retrieval. Diagnostic performance was measured by Macro-F1, whereas retrieval performance was measured by Precision@K, Hit@K, mean reciprocal rank (MRR), and normalized discounted cumulative gain (nDCG@10).

For tomato disease and pest diagnosis, Macro-F1 assigns equal weight to all classes and therefore summarizes diagnostic performance across categories. Precision@K measures the proportion of relevant samples among the top K retrieved candidates and thus reflects the accuracy of a limited result set. For a query q, it is defined as:

$$\mathrm{Precision@}K(q) = \frac{1}{K}\sum_{r=1}^{K}\mathrm{rel}_r(q) = \frac{N_{\mathrm{rel}}^K(q)}{K}, \tag{17}$$

where K is the number of returned results, $\mathrm{rel}_r(q) \in \{0,1\}$ indicates whether the rth candidate for query q is relevant, and $N_{\mathrm{rel}}^K(q)$ is the number of relevant samples among the top-K results. Relevance was determined by whether the query image and candidate image belonged to the same class. A higher Precision@K indicates a greater proportion of relevant samples

among the top K candidates. Overall Precision@K on the test set was calculated by averaging over all queries.

Hit@K indicates whether at least one relevant sample occurs among the top-K results. For a query q, it is defined as:

$$\mathrm{Hit@}K(q) = \begin{cases} 1, & N_{\mathrm{rel}}^{K}(q) > 0, \\ 0, & N_{\mathrm{rel}}^{K}(q) = 0. \end{cases} \tag{18}$$

For a test set containing N_q queries, overall Hit@K is calculated as:

$$\mathrm{Hit@}K = \frac{1}{N_q} \sum_{q=1}^{N_q} \mathrm{Hit@}K(q), \tag{19}$$

Let rank_q denote the rank of the first relevant result for query q. If no relevant result occurs in the retrieval list, $\mathrm{rank}_q = \infty$ and the corresponding reciprocal rank is set to zero. MRR is then defined as:

$$\mathrm{MRR} = \frac{1}{N_q} \sum_{q=1}^{N_q} \frac{1}{\mathrm{rank}_q}, \tag{20}$$

A higher MRR indicates that knowledge items relevant to the query tend to appear earlier in the ranked results. The overall ranking quality of the top 10 results was further evaluated using nDCG@10. For query q,

$$\mathrm{DCG}_q\mathrm{@10} = \sum_{r=1}^{10} \frac{2^{\mathrm{rel}_{q,r}} - 1}{\log_2(r+1)}, \tag{21}$$

where $\mathrm{rel}_{q,r}$ is the graded relevance between query q and the knowledge item at rank r, and $\mathrm{IDCG}_q\mathrm{@10}$ is obtained by ordering the items according to ideal relevance. Thus,

$$\mathrm{nDCG}_q\mathrm{@10} = \frac{\mathrm{DCG}_q\mathrm{@10}}{\mathrm{IDCG}_q\mathrm{@10}}. \tag{22}$$

In the current retrieval experiments, relevance was defined by class agreement. A knowledge item from the same class as the query was considered relevant and assigned a relevance grade of 1; an item from a different class was considered irrelevant and assigned a grade of 0.

2.6 Implementation details

All experiments were conducted on a computing platform equipped with an Intel(R) Xeon(R) Platinum 8270 CPU (2.70 GHz) and an NVIDIA A100-SXM4-80GB GPU. The software environment comprised Ubuntu 22.04 LTS, CUDA 12.1, Python 3.10.14, and PyTorch. Images were resized to 224×224 pixels and normalized using the mean and standard deviation of ImageNet. ResNet-18 was used as the visual encoder, and the learning rate was scheduled using CosineAnnealingLR. Training used class-weighted cross-entropy, label smoothing with a coefficient of 0.05, gradient clipping with a maximum norm of 2.0, and random data augmentation. The training hyperparameters are listed in Table 1.

MRAC was evaluated in controlled synthetic crowdsourcing experiments with 32 simulated annotators and a default of five redundant labels per sample. Annotation noise rates were set to 10%, 20%, 30%, and 40%, and each experiment was repeated independently under five random seeds. For each crowdsourcing seed, approximately 15% of the 3,208 PlantVillage test samples, corresponding to 481 samples, were randomly selected as an annotator-calibration subset and used to estimate the class confusion matrices of the simulated annotators. The remaining 2,727 samples constituted the crowdsourcing evaluation subset for that seed. The calibration subset was excluded from the final metric calculations. Because the calibration and evaluation subsets differed across the five random seeds, crowdsourcing results are reported as the mean $\pm$ standard deviation over five repetitions.

Table 1. Training hyperparameters.

Parameter	Value
Image size	224
Number of classes	10
Batch size	128
Epochs	100
Optimizer	Adam
Learning rate scheduler	CosineAnnealingLR
Weight decay	0.0001
Learning rate	0.0005

3 Experimental Results and Analysis

3.1 Baseline performance of the visual representation

The visual-classification baseline was evaluated on the internal PlantVillage test set to determine whether the resulting representations and class predictions provided stable content priors for subsequent MRAC label aggregation and OMCR knowledge refinement. As shown in Figure 4(a), the cross-entropy losses of the training and validation sets decreased rapidly during early training of the ResNet-18 model with a fine-tuned

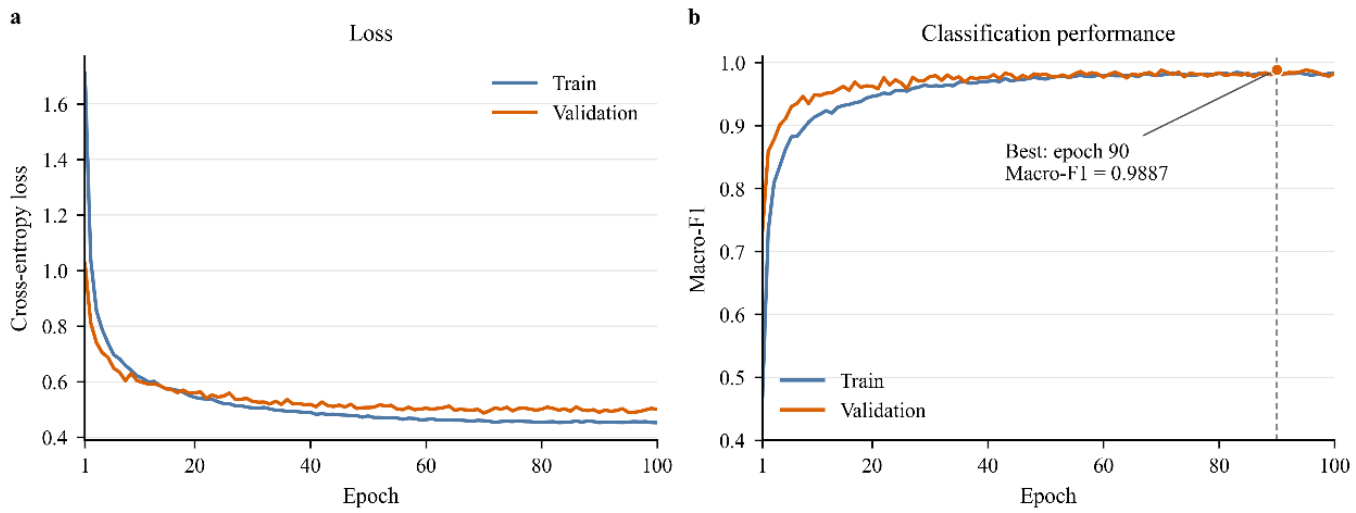


Figure 4. Training dynamics and performance of the visual-classification baseline. (a) Training and validation cross-entropy loss over epochs; (b) training and validation Macro-F1 over epochs.

layer4 block and then gradually stabilized, indicating convergence. Figure 4(b) shows the corresponding changes in training and validation Macro-F1 over the epochs. The maximum validation Macro-F1 was 0.9887; the weights from that epoch were therefore selected as the optimal model. On the internal PlantVillage test set, this model achieved an accuracy of 0.9835, a Macro-F1 of 0.9820, and an ECE of 0.0808. These results indicate balanced classification performance across disease and pest categories. This fine-tuned model served as a reference for visual separability, whereas the calibrated frozen-feature prior described in Section 2.2 supplied the initial content information for the MRAC label posterior and OMCR multi-evidence fusion.

3.2 Evaluation of reliable label aggregation

To evaluate label aggregation under annotation noise and heterogeneous annotator competence, MRAC was compared with majority voting [22], Dawid-Skene [23], a GLAD-inspired method [24], CCC [25], CoNAL [26], and a visual prior without crowdsourcing information (Visual prior only) in a controlled synthetic crowdsourcing environment. The visual-only prior consisted of frozen ResNet-18 features and a multinomial logistic-regression classifier calibrated by temperature scaling. As shown in Table 2, under 30% synthetic annotation noise, MRAC achieved the highest Accuracy, Macro-F1, and Recall@3, with values of 0.9732, 0.9676, and 0.9969, respectively. Relative to CoNAL, MRAC improved Accuracy and Macro-F1 by 0.0105 and 0.0166, respectively; relative to the visual-only prior, the corresponding improvements were 0.0438 and 0.0505. These gains therefore arose not only from the base classifier but also from the joint posterior modeling and updating of annotator reliability, sample difficulty, and crowdsourced evidence. MRAC obtained an ECE of 0.0119, comparable to the 0.0115 achieved by the visual-only prior, indicating that its improved label discrimination did not compromise probability calibration. MRAC thus produced accurate and well-calibrated soft labels under noisy crowdsourcing, which served as the label input for active querying, OMCR multi-evidence clustering, and knowledge refinement.

Sensitivity to random variation was assessed by repeating all methods with the same five fixed random seeds and reporting the mean ± standard deviation. The Accuracy, Macro-F1, and Recall@3 of MRAC were 0.9732 ± 0.0031, 0.9676 ± 0.0044, and 0.9969 ± 0.0009, respectively. The small standard deviations across all three metrics indicate that MRAC was stable across repeated experiments.

As shown in Figure 5, Macro-F1 generally decreased as the annotation noise rate increased. MRAC achieved the highest Macro-F1 at every noise level and exhibited the smallest performance degradation among the methods that use crowdsourced labels. Because the visual-only prior did not use crowdsourced labels, its performance remained essentially stable across noise rates, but was consistently lower than that of MRAC. MRAC therefore benefits from crowdsourced information while limiting the adverse effects of highly noisy annotations.

Table 2. Performance comparison of crowdsourced label-aggregation methods under 30% annotation noise.

Model	Accuracy	Macro-F1	Recall@3	ECE
Majority vote	0.8728 ± 0.0030	0.8429 ± 0.0024	0.9718 ± 0.0032	0.2320 ± 0.0021
Dawid-Skene	0.8748 ± 0.0038	0.8349 ± 0.0064	0.9687 ± 0.0031	0.0727 ± 0.0049
GLAD-inspired	0.8828 ± 0.0048	0.8523 ± 0.0066	0.9702 ± 0.0031	0.0227 ± 0.0071
CCC	0.9498 ± 0.0045	0.9352 ± 0.0067	0.9917 ± 0.0016	0.0226 ± 0.0024
CoNAL	0.9627 ± 0.0052	0.9510 ± 0.0078	0.9940 ± 0.0017	0.0197 ± 0.0033
Visual prior only	0.9294 ± 0.0024	0.9171 ± 0.0040	0.9937 ± 0.0004	0.0115 ± 0.0012
MRAC	0.9732 ± 0.0031	0.9676 ± 0.0044	0.9969 ± 0.0009	0.0119 ± 0.0017

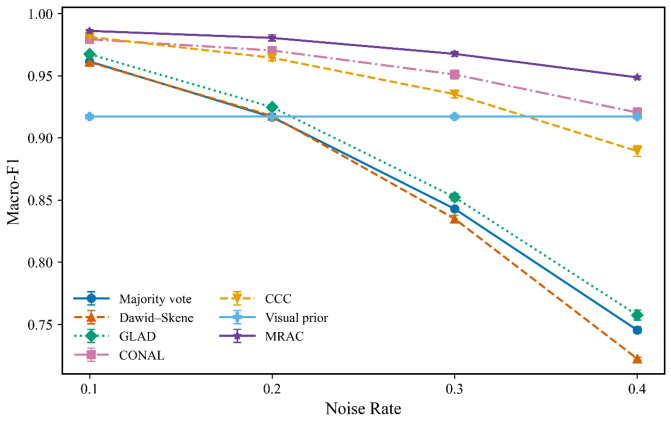


Figure 5. Robustness comparison of crowdsourced label-aggregation methods under different annotation noise rates.

3.3 Cost efficiency of active querying

The efficiency of MRAC-EIG under a limited annotation budget was evaluated at 30% synthetic annotation noise by comparison with random querying and uncertainty-based querying while increasing the maximum number of queries per sample from one to five. As shown in Table 3, when up to five queries were permitted, MRAC-EIG achieved the highest Macro-F1 (0.9631 ± 0.0087), with a mean of 1.2189 ± 0.0137 queries and a mean annotation cost of 1.0080 ± 0.0353. Random querying achieved a Macro-F1 of 0.9591 ± 0.0041, with 1.2251 ± 0.0120 queries and a mean cost of 1.3036 ± 0.0363. Uncertainty-based querying achieved a Macro-F1 of 0.9557 ± 0.0081, with 1.1997 ± 0.0307 queries and a mean cost of 1.2502 ± 0.0515. Thus, the performance gain of MRAC-EIG did not result from issuing substantially more queries. Instead, the method jointly considered class-posterior uncertainty, annotator reliability, and per-query cost to allocate limited resources to informative sample-annotator pairs. In this controlled setting, MRAC-EIG therefore reduced total annotation cost while maintaining strong label-aggregation performance. As shown in Figure 6, Macro-F1 generally increased with the query budget for all three strategies, whereas MRAC-EIG achieved higher classification performance at a lower mean annotation cost. The observed mean number of queries increased only to approximately 1.20-1.23, indicating that the joint stopping mechanism effectively eliminated unnecessary repeated queries.

3.4 Evaluation of multi-evidence clustering performance

A progressive ablation study was conducted to quantify the contributions of different evidence sources to OMCR clustering and knowledge refinement. Crowdsourced soft posteriors, ontology constraints, and the complete refinement mechanism were added sequentially to the content features, as shown in Table 4. With content features alone, ACC, NMI, and ARI were 0.4953, 0.4867, and 0.3785, respectively, indicating that a single content representation provided limited separation of the class structure. Adding MRAC soft posteriors increased these metrics to 0.9704, 0.9285, and 0.9494, respectively, while Macro-F1 increased from 0.9252 to 0.9654. Reliable crowdsourcing posteriors therefore provided the largest improvement in class alignment and clustering consistency. The subsequent inclusion of ontology constraints further increased ACC, NMI, and ARI to 0.9707, 0.9293, and 0.9499, respectively. The complete OMCR model achieved the highest ACC, ARI, AUPRC, and Macro-F1, with values of 0.9708, 0.9500, 0.9977, and 0.9656, respectively; its NMI was 0.9293, equal to the value obtained after adding ontology constraints. Multi-evidence fusion thus substantially improved the cluster structure relative to content features alone. MRAC soft posteriors contributed most of the gain, while ontology constraints and the subsequent refinement mechanism marginally improved clustering consistency and incorrect-label detection.

Table 3. Performance and annotation efficiency of different active querying policies with a maximum of five queries per sample.

Query policy	Macro-F1	Mean queries	Cost
Random querying	0.9591 ± 0.0041	1.2251 ± 0.0120	1.3036 ± 0.0363
Uncertainty-based querying	0.9557 ± 0.0081	1.1997 ± 0.0307	1.2502 ± 0.0515
MRAC-EIG	0.9631 ± 0.0087	1.2189 ± 0.0137	1.0080 ± 0.0353

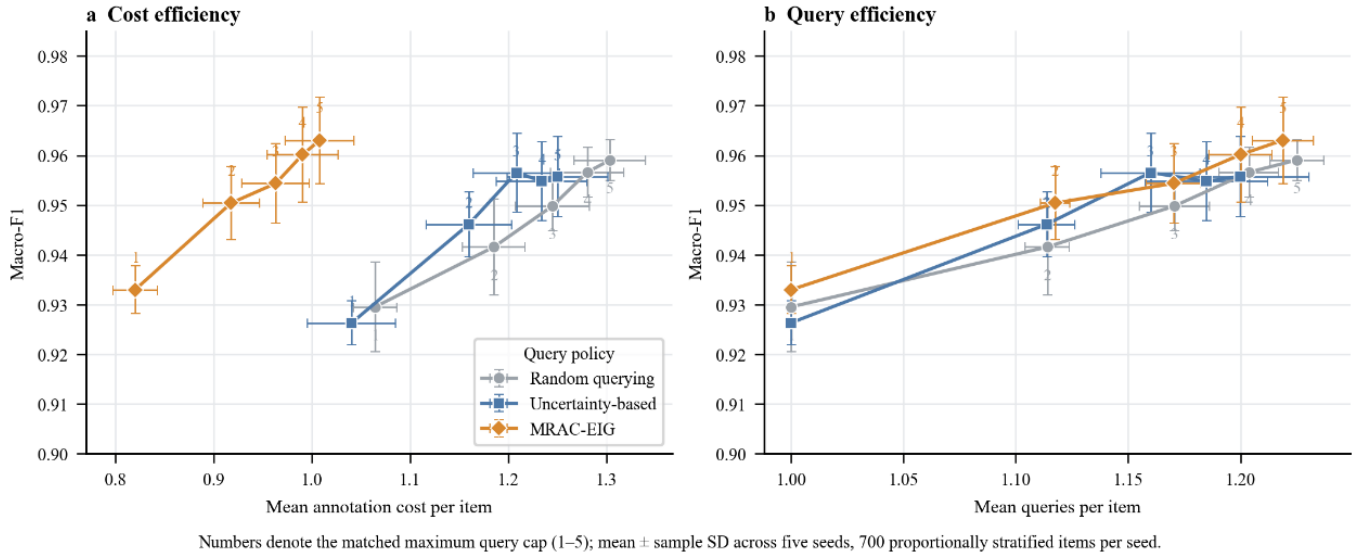


Figure 6. Efficiency of active-query policies under different maximum query limits. (a) Relationship between Macro-F1 and mean annotation cost per item; (b) relationship between Macro-F1 and mean number of queries per item.

Table 4. Progressive ablation results for OMCR evidence sources and refinement components.

Methods	ACC	NMI	ARI	AUPRC	Macro-F1
Content only	0.4953	0.4867	0.3785	0.9905	0.9252
+MRAC	0.9704	0.9285	0.9494	0.9957	0.9654
+Ontology	0.9707	0.9293	0.9499	0.9957	0.9654
OMCR	0.9708	0.9293	0.9500	0.9977	0.9656

3.5 Retrieval of class-related image instances

This experiment assessed the effects of MRAC crowdsourcing posteriors and OMCR label refinement on agricultural image retrieval. A fixed set of 700 queries was sampled from the PlantVillage test set and reused for every random seed; the top 10 candidates were retrieved from the PlantVillage training set. As shown in Table 5, visual-only retrieval achieved Precision@5, Precision@10, and nDCG@10 values of 0.8151, 0.7906, and 0.8031, respectively. With MRAC posteriors alone (MRAC only), the corresponding metrics increased to 0.9202, 0.9138, and 0.9171, indicating that the MRAC posterior supplied useful class-semantic information for retrieval. With Visual + MRAC + OMCR, Precision@5, Precision@10, and nDCG@10 further increased to 0.9251, 0.9223, and 0.9233, respectively; nDCG@10 improved by approximately 15.0% relative to visual-only retrieval. Compared with MRAC alone, Visual + MRAC + OMCR improved Precision@5, Precision@10, and nDCG@10; Hit@5 was essentially unchanged (0.9514 vs. 0.9511), whereas Hit@10 and MRR decreased slightly. OMCR therefore improved candidate-set purity and overall ranking quality rather than the rank of the first relevant sample or the probability of retrieving at least one relevant sample.

Figures 7(a) and 7(c) show that introducing MRAC improved both overall ranking quality and top-K retrieval precision, while OMCR refinement produced additional gains in nDCG@10, Precision@5, and Precision@10. Figure 7(b) shows that Hit@5 and Hit@10 remained above 0.95 for all methods, with visual-only retrieval giving the highest values. Figure 7(d) indicates that the per-query latency of Visual + MRAC + OMCR was 0.1527 ms, compared with 0.1220 ms for visual-only retrieval, representing

Table 5. Comparison of image retrieval performance under different retrieval strategies.

Retrieval method	Precision@5	Precision@10	Hit@5	Hit@10	MRR	nDCG@10	Latency (ms)
Visual only	0.8151	0.7906	0.9643	0.9771	0.9019	0.8031	0.1220
MRAC only	0.9202	0.9138	0.9511	0.9557	0.9400	0.9171	0.1313
Visual + MRAC	0.9201	0.9136	0.9511	0.9557	0.9400	0.9169	0.1568
Visual + MRAC + OMCR	0.9251	0.9223	0.9514	0.9528	0.9378	0.9233	0.1527

a small additional delay. The MRAC posterior was thus the principal source of the retrieval improvement, whereas OMCR further optimized candidate ranking and label consistency. In the controlled PlantVillage setting, the proposed approach improved the retrieval and organization of same-class samples in the knowledge base and can support searches for similar cases and the selection of candidate evidence.

4 Discussion

This study developed a unified framework comprising MRAC and OMCR to address unstable annotation quality, the limited integration of heterogeneous evidence, and unreliable knowledge organization in agricultural multimedia knowledge construction. Incorporating annotator reliability, sample difficulty, multimodal evidence quality, and annotation cost into a unified posterior model improved crowdsourced label aggregation under controlled noise. At 30% synthetic annotation noise, MRAC achieved an accuracy of 0.9732 and a Macro-F1 of 0.9676, exceeding CoNAL by 0.0105 and 0.0166, respectively, while maintaining an ECE of 0.0119. Thus, the performance gain did not incur an appreciable loss of probability calibration. The active-query experiments further demonstrated the cost efficiency of MRAC-EIG. With comparable mean query counts, MRAC-EIG achieved a Macro-F1 of 0.9631 at a mean relative annotation cost of 1.0080. Its mean cost was approximately 22.7% lower than that of random querying and 19.4% lower than that of uncertainty-based querying. Because all three strategies issued approximately 1.20 queries per sample on average, the cost advantage of MRAC-EIG arose primarily from the joint selection of samples, annotators, and per-query costs rather than from an increase in repeated annotations. These findings support the ability of MRAC-EIG to improve the use of annotation resources in the present controlled experiments.

Existing approaches to multimedia knowledge construction generally address budget allocation, repeated-annotation selection, and annotator-reliability estimation separately [13, 25]. In contrast, MRAC incorporates sample uncertainty, annotator confusion patterns, modality quality, and annotation cost into query decisions at the sample-annotator level. Multiview clustering methods such as MFLVC and ICMVC primarily improve representation learning through cross-view consistency and high-confidence guidance [14, 15], but they do not generally integrate crowdsourcing posteriors and agricultural ontology constraints into the same clustering and risk-routing process. Agri-LLaVA, CDDM, and TomaMMU focus on agricultural knowledge understanding and diagnostic reasoning using curated image-text or visual question-answering data [7, 8, 10], whereas CropDP-KG relies on structured knowledge subjected to manual annotation and expert quality control [17]. The present study instead addresses reliable annotation, noise detection, and semantic refinement before such data enter a knowledge system, linking crowdsourced annotation, clustering analysis, and retrieval evaluation into a continuous workflow. In this way, the uncertainty estimated at the annotation stage is carried forward to clustering and retrieval rather than discarded.

This study nevertheless has several limitations. MRAC was evaluated in controlled experiments with 32 simulated annotators and predefined noise rates, and the relative annotation costs were also generated synthetically. Future work should conduct real crowdsourcing experiments involving agricultural experts, extension personnel, and nonexpert annotators, while recording annotation time, actual expenditure, and behavioral changes and using statistical tests to compare query strategies. In addition, the principal quantitative experiments used relatively controlled datasets such as PlantVillage. The evidence views did not constitute a large-scale collection of strictly paired images, text, and videos at the instance level; the results therefore cannot fully establish performance in real cross-modal settings. Future datasets should include natural field samples collected across regions, seasons, and cultivars, together with paired images, videos, textual

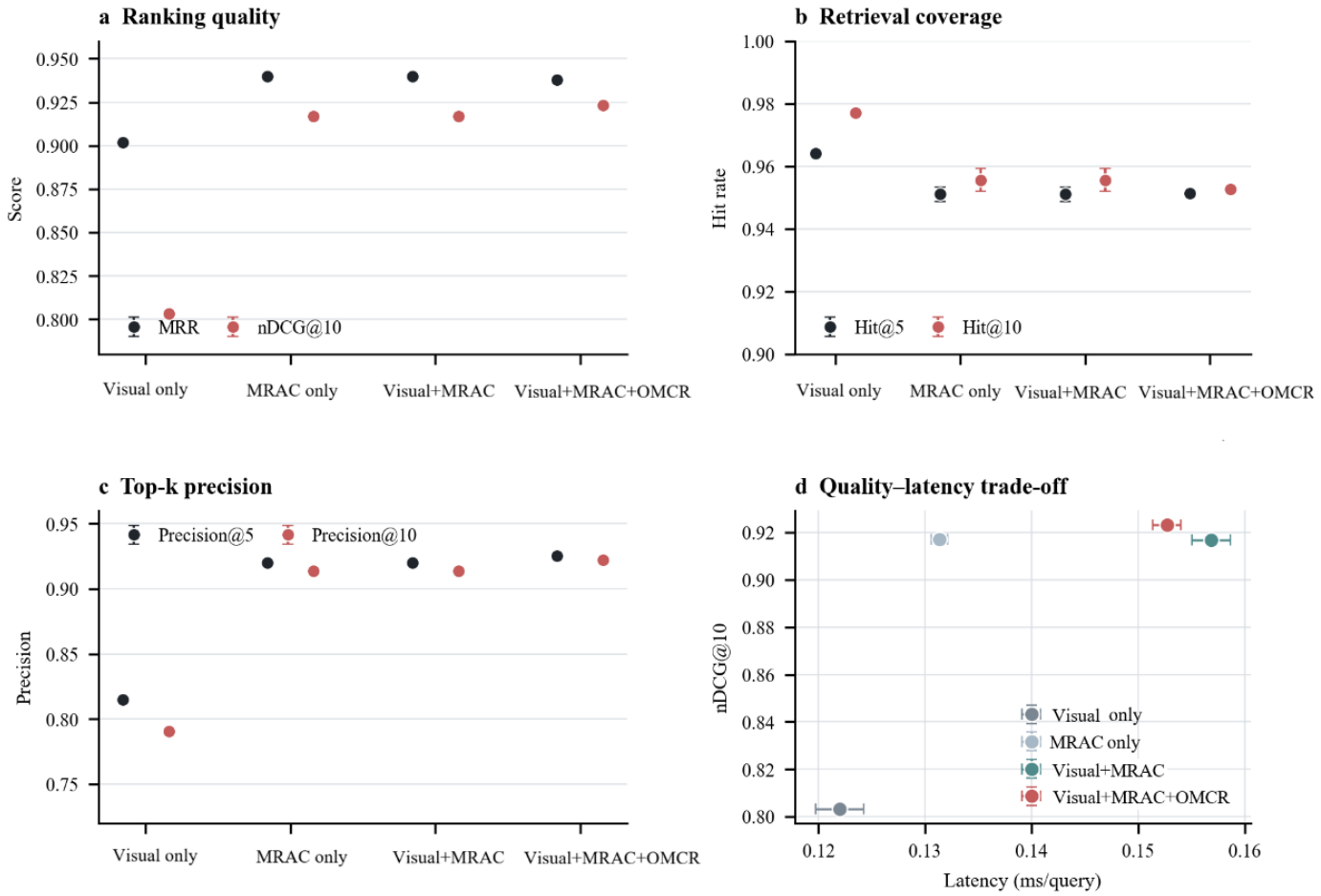


Figure 7. Comparison of retrieval quality and computational efficiency among different image retrieval strategies. (a) Ranking quality; (b) retrieval coverage; (c) top-K precision; (d) quality-latency trade-off.

descriptions, and ontology knowledge for the same disease event.

5 Conclusion

This study proposed a unified approach to agricultural multimedia knowledge construction comprising MRAC and OMCR. MRAC improved crowdsourced label quality through reliability modeling and active querying while accounting for annotation cost. OMCR integrated content features, crowdsourcing posteriors, and agricultural ontology information to support multimedia clustering, incorrect-label detection, and knowledge refinement. In the controlled experiments, MRAC improved noisy-label aggregation, whereas OMCR increased multi-evidence clustering consistency and the ranking quality of class-related image retrieval. Generalization to real crowdsourcing, instance-level paired multimodal data, and natural field environments remains to be evaluated. The framework offers a practical route for converting heterogeneous agricultural data into structured and trustworthy knowledge and can support the construction and retrieval of agricultural knowledge bases.

Data Availability Statement

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Funding

This work was supported by the National Major Agricultural Science and Technology Project under Grant 0304.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in

the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., & Bochtis, D. (2021). Machine learning in agriculture: A comprehensive updated review. *Sensors, 21*(11), 3758. [CrossRef]

[2] Lv, B., Wu, H., Chen, W., Chen, C., Miao, Y., & Zhao, C. (2024). VEG-MMKG: Multimodal knowledge graph construction for vegetables based on pre-trained model extraction. *Computers and electronics in agriculture, 226*, 109398. [CrossRef]

[3] Liu, X., Bai, X., Wang, L., Ren, B., Lu, S., & Lin, L. (2019). Review and trend analysis of knowledge graphs for crop pest and diseases. *IEEE Access, 7*, 62251-62264. [CrossRef]

[4] Zhao, X., Chen, B., Ji, M., Wang, X., Yan, Y., Zhang, J., ... & Lv, C. (2024). Implementation of large language models and agricultural knowledge graphs for efficient plant disease detection. *Agriculture, 14*(8), 1359. [CrossRef]

[5] Dolatabadian, A., Neik, T. X., Danilevicz, M. F., Upadhyaya, S. R., Batley, J., & Edwards, D. (2025). Image-based crop disease detection using machine learning. *Plant Pathology, 74*(1), 18-38. [CrossRef]

[6] Zhu, H., Qin, S., Su, M., Lin, C., Li, A., & Gao, J. (2025). Harnessing large vision and language models in agriculture: A review. *Frontiers in Plant Science, 16*, 1579355. [CrossRef]

[7] Wang, L., Jin, T., Yang, J., Leonardis, A., Wang, F., & Zheng, F. (2024). Agri-llava: Knowledge-infused large multimodal assistant on agricultural pests and diseases. *arXiv preprint arXiv:2412.02158*. [CrossRef]

[8] Liu, X., Liu, Z., Hu, H., Chen, Z., Wang, K., Wang, K., & Lian, S. (2024, September). A multimodal benchmark dataset and model for crop disease diagnosis. In *European conference on computer vision* (pp. 157-170). Cham: Springer Nature Switzerland. [CrossRef]

[9] Zhang, M., Xu, Z., Wang, P., Li, R., Wang, L., Liu, Q., ... & Wu, S. (2026, May). Agridoctor: A multimodal intelligent assistant for agriculture. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2741-2745). IEEE. [CrossRef]

[10] Truong, G. H., Quoc, K. N., & Quach, L. D. (2026). TomaMMU: A Comprehensive Multimodal Understanding Benchmark for Tomato Leaf Diseases. *arXiv preprint arXiv:2608.08727*. [CrossRef]

[11] Zhang, J., Wu, M., Zhou, C., & Sheng, V. S. (2022). Active crowdsourcing for multilabel annotation. *IEEE Transactions on Neural Networks and Learning Systems, 35*(3), 3549-3559. [CrossRef]

[12] Majdi, M. S., & Rodriguez, J. J. (2023). Crowd-Certain: label aggregation in crowdsourced and ensemble learning classification. *arXiv preprint arXiv:2310.16293*. [CrossRef]

[13] Goh, H. W., & Mueller, J. (2023). Activelab: Active learning with re-labeling by multiple annotators. *arXiv preprint arXiv:2301.11856*. [CrossRef]

[14] Xu, J., Tang, H., Ren, Y., Peng, L., Zhu, X., & He, L. (2022, June). Multi-level feature learning for contrastive multi-view clustering. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 16030-16039). IEEE. [CrossRef]

[15] Chao, G., Jiang, Y., & Chu, D. (2024, March). Incomplete contrastive multi-view clustering with high-confidence guiding. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 38, No. 10, pp. 11221-11229). [CrossRef]

[16] Wang, K., Miao, Y., Wang, X., Li, Y., Li, F., & Song, H. (2024). Research on the construction of a knowledge graph for tomato leaf pests and diseases based on the named entity recognition model. *Frontiers in Plant Science, 15*, 1482275. [CrossRef]

[17] Yan, R., An, P., Meng, X., Li, Y., Li, D., Xu, F., & Dang, D. (2025). A knowledge graph for crop diseases and pests in China. *Scientific Data, 12*, 222. [CrossRef]

[18] Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in plant science, 7*, 215232. [CrossRef]

[19] Cooper, L., Walls, R. L., Elser, J., Gandolfo, M. A., Stevenson, D. W., Smith, B., Preece, J., Athreya, B., Mungall, C. J., & Rensing, S. (2013). The plant ontology as a tool for comparative plant anatomy and genomic analyses. *Plant and Cell Physiology, 54*(2), e1. [CrossRef]

[20] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778). [CrossRef]

[21] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1321-1330). PMLR. http://proceedings.mlr.press/v70/guo17a.html

[22] Chen, Z., Jiang, L., & Li, C. (2022). Label augmented and weighted majority voting for crowdsourcing. *Information Sciences, 606*, 397-409. [CrossRef]

[23] Dawid, A. P., & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 28*(1), 20-28. [CrossRef]

[24] Whitehill, J., Wu, T., Bergsma, J., Movellan, J., & Ruvolo, P. (2009). Whose vote should count

more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, *22*.

[25] Zhang, H., Li, S., Zeng, D., Yan, C., & Ge, S. (2024). Coupled confusion correction: Learning from crowds with sparse annotations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(15), 16732-16740. [CrossRef]

[26] Chu, Z., Ma, J., & Wang, H. (2021). Learning from crowds by modeling common confusions. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(7), 5832-5840. [CrossRef]



Hongda Li is a Ph.D. candidate in Agricultural Electrification and Automation at the College of Information and Electrical Engineering, China Agricultural University, Beijing, China. He is also a jointly supervised doctoral student at the Beijing Academy of Agriculture and Forestry Sciences, Beijing, China. His research focuses on intelligent agricultural knowledge services, covering the complete process from data perception to intelligent decision-making in smart agriculture. (Email: lihd@nercita.org.cn)



Zixin Zhu is a master's student in Agricultural Engineering and Information Technology at the College of Information Science and Engineering, Shandong Agricultural University. She is currently involved in research with the Big Data Intelligence Team at the Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences. Her work focuses on the development and application of intelligent algorithms for agriculture and knowledge-driven foundation models. (Email: 2025121221@sdau.edu.cn)



Guanxu Ren is a master's student in Control Science and Engineering at the School of Automation, Beijing Information Science and Technology University (enrolled in 2024). His research primarily focuses on machine vision, natural language processing, and error correction mechanisms for large models. (Email: 2024020398@bistu.edu.cn)



Yisheng Miao received the Ph.D. degree and is a Professor. He currently serves as an expert of the Beijing Smart Agriculture Innovation Team, a member of the intelligent management team of the National Bulk Vegetable Industry Technology System, and an expert of the "Innovation China" Modern Agriculture Industry Service Group. He has long been engaged in the research, development, and application of agricultural knowledge-driven large models and autonomous unmanned operation systems. He developed the agricultural multimodal large model "Qiji" with a knowledge base of tens of millions of entries, and has worked on intelligent operation integrating agronomy and agricultural machinery, reliable sensing optimization for the Internet of Things, and agricultural domain-specific large models. (Email: miaoys007@nercita.org.cn)



Peng Hao received the Bachelor's degree and is an Associate Senior Agronomist. He has long been engaged in demonstration projects of the agricultural Internet of Things, technical research on agricultural socialized services, smart agriculture, and the promotion and demonstration of agricultural technologies. (Email: haop007@nercita.org.cn)