Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

GEPC: Generative-Explanatory Pest Classification with Multi-Modal Priors and Hierarchical Concept Constraints

Yuting Xian
Sichuan Energy Internet Research Institute, Tsinghua University
Yupei Song
songyupei@tsinghua-eiri.org
Sichuan Energy Internet Research Institute, Tsinghua University

Research Article

Keywords: Pest Classification, Multimodal Fusion, Hierarchical Sparse Coding, Interpretability, Long-Tailed Recognition

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11578453/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

GEPC: Generative-Explanatory Pest Classification with Multi-Modal Priors and Hierarchical Concept Constraints

Yuting Xian[1] **and Yupei Song**[1,*]

[1]Sichuan Energy Internet Research Institute, Tsinghua University, Chengdu, 610213, China
[*]Corresponding author: Yupei Song (e-mail: songyupei@tsinghua-eiri.org)

ABSTRACT

Accurate and interpretable pest classification underpins precision agriculture and food security, yet three obstacles remain. Field-collected datasets are severely long-tailed, so rare and often high-risk pests are under-represented; closely related species are morphologically near-identical, limiting single-modality discrimination; and standard end-to-end classifiers emit labels without auditable reasoning. We present GEPC (Generative-Explanatory Pest Classifier), a unified framework organized around one principle: every prediction should be backed by a verifiable taxonomic pathway. To counter the long tail, a semantics-guided diffusion pipeline synthesizes ecologically plausible samples for rare classes through structured prompts and CLIP-based quality filtering, grounding augmentation in real taxonomic and habitat priors. For interpretability and fine-grained discrimination, GEPC casts feature extraction as non-negative, path-constrained sparse pursuit via Hierarchical Beam Orthogonal Matching Pursuit (HB-OMP) over an offline differential dictionary built on the Order–Family–Species tree, so the recovered root-to-leaf path denoises a frozen DINOv2 visual representation while doubling as a transparent diagnostic cue. This hierarchy-aligned signal is fused with species-anonymized morphological descriptions encoded by a frozen CLIP text encoder, and a lightweight classifier issues the final prediction. Across multiple random seeds on AP162, GEPC attains $90.4 \pm 0.3\%$ overall accuracy, 82.4% tail-class recall, and 94.6% agreement between the visual pursuit path and the final decision, consistently outperforming end-to-end supervised and zero-shot baselines while exposing the taxonomic evidence behind each prediction.
Keywords: Pest Classification; Multimodal Fusion; Hierarchical Sparse Coding; Interpretability; Long-Tailed Recognition

Introduction

Accurate agricultural pest classification is a cornerstone of crop pest management and food security, with the Food and Agriculture Organization (FAO) estimating that pest damage causes 20%–40% of global annual crop losses[1]. Nevertheless, current methods face three core bottlenecks: the inherent long-tailed data distribution, where high-risk rare pests have extremely limited samples, severely weakening model performance on minority classes[2–4]; the high morphological similarity between closely related species, where fine-grained discriminative features are difficult to extract via a single visual modality, restricting classification accuracy[5]; and the opaque decision-making process of black-box models, which output category labels without verifiable reasoning chains[6].

Existing approaches mainly fall into two paradigms, as shown in Fig. 1. The first employs end-to-end trained deep visual models[7] such as ResNet[8] and Vision Transformer[9], which deliver competitive performance on balanced benchmarks but perform poorly on few-shot fine-grained pest categories and remain black-box systems without interpretability. The second applies pre-trained vision-language models such as CLIP[10] and multimodal large language models for zero-shot classification[11]; however, their general-domain pre-training lacks adaptation to pest-specific features, leading to suboptimal recognition accuracy.

To address these bottlenecks, we argue that three complementary strategies can be effectively combined. First, generative data augmentation can synthesize high-fidelity, ecologically plausible samples for rare pest categories, directly mitigating the long-tailed class imbalance. Second, pre-trained vision-language models serve as robust cross-modal encoders that infuse domain semantic knowledge, compensating for the fine-grained visual ambiguities among morphologically similar species. Third, the canonical insect taxonomic hierarchy inherently encodes rich morphological and ecological priors, providing a natural scaffold for constructing transparent, verifiable reasoning chains.

Motivated by the above analysis, we propose GEPC (Generative-Explanatory Pest Classifier), a unified framework that tackles the three core bottlenecks through three tightly integrated components, as illustrated in Fig. 2. First, a semantics-guided diffusion pipeline synthesizes diverse, ecologically plausible samples for rare pest categories. Second, a hierarchical sparse coding dictionary is pre-built over the Order–Family–Species taxonomic tree, and a dedicated Hierarchical Beam Orthogonal Matching Pursuit (HB-OMP)[12] solver constrains visual feature extraction to valid taxonomic paths, thereby converting opaque

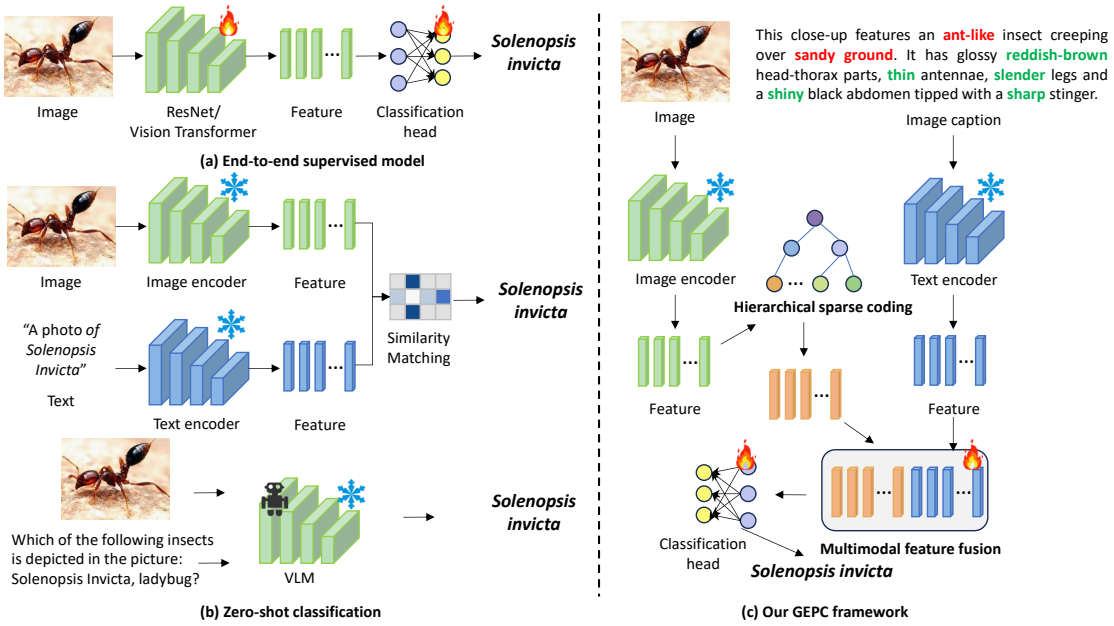


Figure 1. Overview of existing methods and our GEPC framework.

visual representations into transparent, verifiable hierarchical trajectories. Third, a dual-modal alignment module fuses frozen DINOv2[13] visual features with CLIP-encoded morphological descriptions, substantially enhancing fine-grained discriminability among morphologically similar species. These components operate under an offline–online paradigm: the diffusion-based augmentation, dictionary construction, and training-set text description generation are performed offline once, while the online phase executes hierarchical pursuit, multi-modal fusion, and final species prediction with all foundation models frozen.

In summary, the main contributions of this paper are as follows:

- We propose a semantics-guided diffusion augmentation pipeline with structured prompt construction and CLIP-based quality filtering, which synthesizes high-fidelity samples for rare pest categories exclusively from training-split metadata and effectively alleviates the long-tailed class imbalance problem.
- We adapt parent–child differential sparse coding to the biological Order–Family–Species taxonomy by introducing positive directional correlation and non-negative least-squares constraints into fixed-depth HB-OMP decoding. This resolves sibling collinearity and single-child atom degeneracies while guaranteeing that every visual pursuit terminates at a complete root-to-leaf path.
- We design a dual-modal fusion architecture that aligns hierarchy-reconstructed visual features with species-anonymized morphological descriptions via a lightweight learnable projection head, substantially enhancing feature discriminability for fine-grained pest identification.
- Extensive experiments on the AP162 dataset show that GEPC achieves 90.4% overall accuracy (82.4% tail-class recall and 94.6% path–classifier agreement), outperforming matched-backbone and hierarchical baselines while degrading gracefully to reliable family- and order-level predictions when species-level visual cues are ambiguous.

Related Work

Agricultural Pest Classification. Traditionally, pest recognition has relied heavily on human experts, a process that is costly, time-consuming, and impractical for large-scale agricultural settings. Early automated pest recognition relied on hand-crafted features paired with traditional classifiers such as SVMs[14,15]. The prevalence of deep learning, propelled by large-scale benchmarks including IP102[16] and AP162[17], has substantially boosted classification performance via backbone architecture iteration and data augmentation strategies[18]. Nevertheless, most existing works frame pest classification as a standard end-to-end visual task, largely overlooking the severe long-tailed class imbalance in field-collected datasets and the indispensable demand for interpretable decisions for credible agricultural management[18,19]. Different from prior black-box paradigms, our work reformulates visual feature extraction as a constrained hierarchical path-discovery problem by embedding taxonomic structures

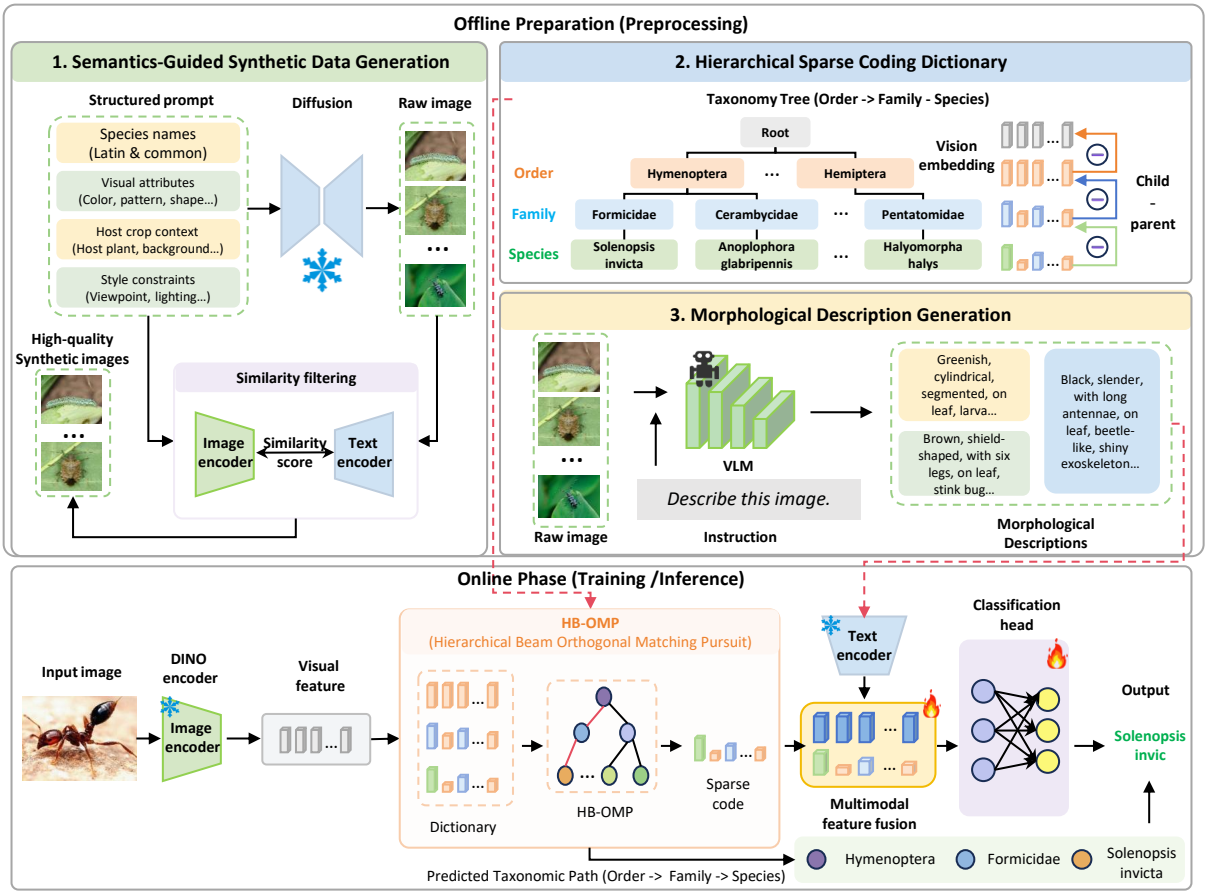


Figure 2. Overall framework of GEPC operating across offline preparation and online inference phases.

into the representation space; instead of conventional oversampling[20] or loss re-weighting[21], we adopt a generative synthesis pipeline to enrich feature diversity of minority classes, achieving both improved rare-category performance and transparent taxonomic reasoning.

Model Interpretability in Visual Recognition. Visual interpretability research falls into two main categories: post-hoc explanation methods (e.g., Grad-CAM[22], SHAP[23]) and inherently interpretable paradigms. As a representative sparse coding-based interpretable approach, HCEP[12] introduces parent–child differential dictionaries and hierarchical beam pursuit (HB-OMP) to realize path-wise interpretable classification, breaking the limitation of flat concept organization in traditional sparse coding methods. However, HCEP[12] relies solely on unimodal visual features and unrestricted least-squares pursuit, which suffers from sign-flipping collinearity in two-species families and lacks mechanisms to handle long-tailed class distributions or fine-grained morphological overlap. In contrast, our GEPC framework extends hierarchical sparse coding by: (1) enforcing positive directional atom selection and non-negative coefficient constraints to eliminate sibling path ambiguity in biological taxonomies; (2) coupling hierarchy-reconstructed visual representations with VLM-derived morphological text embeddings; and (3) integrating semantics-guided diffusion augmentation to stabilize dictionary atom estimation and classifier decision boundaries on tail categories.

Methodology

Overall Framework

The overall architecture of GEPC is illustrated in Fig. 2. The framework operates in two phases: an offline preparation phase and an online training/inference phase. In the offline phase, three independent preprocessing steps are performed strictly on the training split: (i) a semantics-guided diffusion pipeline synthesizes high-fidelity samples for rare pest categories, with CLIP-based quality filtering and human verification to ensure label fidelity; (ii) a hierarchical sparse coding dictionary $\mathbf{D}$ is constructed over the Order–Family–Species taxonomic tree using real training images, where each atom encodes the visual difference between a child node and its parent, enabling telescoping path representations; and (iii) a vision-language model

generates fine-grained, species-anonymized morphological descriptions, which are encoded by a frozen CLIP text encoder and stored for training fusion.

In the online phase, all foundation models remain frozen. For an input image $\mathbf{I}$, a frozen DINOv2 encoder extracts a visual feature $\mathbf{f}_{\text{vis}}$, which is passed to the HB-OMP solver. HB-OMP produces a sparse coefficient vector α whose support strictly follows a complete root-to-leaf taxonomic path $\mathscr{S}$. The reconstructed, hierarchy-aligned visual feature $\mathbf{f}_{\text{vis}}^{\text{rec}} = \mathbf{D}\alpha$ is then concatenated with the CLIP text embedding $\mathbf{f}_{\text{txt}}$, and a lightweight learnable projection head followed by a linear classifier produces the final species prediction $\hat{\mathbf{y}}$. During inference, GEPC outputs both the multimodal species prediction $\hat{\mathbf{y}}$ and the visual taxonomic path $\mathscr{S}$ recovered by HB-OMP, enabling practitioners to inspect the hierarchical visual evidence and verify its consistency with the final decision.

Semantics-Guided Synthetic Data Generation

We augment minority classes through a three-stage pipeline: structured prompt construction, diffusion-based synthesis, and CLIP-based quality filtering. Given a training set with per-class sample counts n_c, we define a target count $N_{\text{target}} = 1{,}500$ and, for each minority class with $n_c < N_{\text{target}}$, synthesize candidate images to narrow the sample gap, producing 96,400 raw candidates in total (Table 3). To prevent data leakage, all prompt construction, image synthesis, and quality filtering are conducted exclusively on the training split; validation and test splits are never accessed.

Structured Prompt Construction. To avoid semantic ambiguity, we design prompts from four complementary components:

$$\begin{aligned}\texttt{Prompt} = {} & [\textit{Species Name}] + [\textit{Visual Attributes}] \\ & + [\textit{Host Crop Context}] + [\textit{Style Constraints}],\end{aligned} \tag{1}$$

where the Latin binomial ensures taxonomic precision, visual attributes encode morphological traits (e.g., body color, wing venation), host crop context specifies ecologically plausible backgrounds, and style constraints introduce variations in lighting, viewpoint, and pose. Species names and host crops are obtained from the AP162 training metadata, while morphological attributes are derived from entomological keys. Multiple prompt variants are generated per species to cover diverse visual appearances.

Diffusion-Based Synthesis. We adopt Stable Diffusion v2.1[24] as the generative backbone. For each minority class, the constructed prompts are fed into the diffusion model to synthesize candidate images conditioned on species identity and ecological context.

CLIP-Based Quality Filtering and Manual Screening. Synthetic images may suffer from text–image misalignment (e.g., missing subjects or morphological distortions). We first filter out candidate samples whose CLIP cosine similarity between the generated image and its prompt falls below a threshold $\tau_{\text{CLIP}} = 0.30$. We then manually screen the retained images and discard those exhibiting severe anatomical artifacts (e.g., fused wings, detached appendages, or absent insect bodies); the number of images retained at each stage is reported in Table 3.

Hierarchical Sparse Coding Dictionary Construction

To enable interpretable hierarchical reasoning, we construct a sparse coding dictionary $\mathbf{D}$ over the Order–Family–Species taxonomic tree. Each atom encodes the visual difference between a child node and its parent, so that any root-to-leaf path corresponds to a telescoping sparse combination.

Taxonomic Hierarchy and Independent Verification. Since AP162 provides only species-level labels, we initially queried GPT-4[25] to draft Order and Family annotations for the 162 species, and subsequently cross-checked and verified every species–family–order mapping against the authoritative EPPO Global Database[26] and standard entomological taxonomy. The verified three-level tree (excluding the virtual root) comprises 9 Orders, 76 Families, and 162 Species leaves, totaling $|\mathscr{N}| = 247$ non-root nodes.

Node Embeddings. All dictionary embeddings are computed exclusively from real training images (excluding synthetic images to preserve authentic taxonomic prototypes). A frozen DINOv2 encoder extracts patch-level features, which are max-pooled into a global visual feature $\mathbf{f}_{\text{vis}} \in \mathbb{R}^{D_v}$. For each leaf i, its embedding $\mathbf{a}^{(i)}$ is the average $\mathbf{f}_{\text{vis}}$ over all its real training images; for an internal node, $\mathbf{a}^{(i)}$ is the average of its children's embeddings:

$$\mathbf{a}^{(i)} = \begin{cases} \frac{1}{|\mathscr{S}_i|} \sum_{\mathbf{I} \in \mathscr{S}_i} \mathbf{f}_{\text{vis}}(\mathbf{I}), & \text{if } i \text{ is a leaf}, \\ \frac{1}{|\text{chi}(i)|} \sum_{j \in \text{chi}(i)} \mathbf{a}^{(j)}, & \text{otherwise}, \end{cases} \tag{2}$$

where $\mathscr{S}_i$ denotes the set of real training images of species i and $\text{chi}(i)$ the set of children of node i.

Differential Dictionary and Degenerate Branch Handling. For every non-root node i, the dictionary atom is defined as $\mathbf{d}^{(i)} = \mathbf{a}^{(i)} - \mathbf{a}^{(\mathrm{par}(i))}$, with the root embedding set to $\mathbf{0}$. Stacking all atoms yields $\mathbf{D} = [\mathbf{d}^{(i)}]_{i \in \mathscr{N}} \in \mathbb{R}^{D_v \times |\mathscr{N}|}$. By this telescoping construction, the embedding of any node i can be expressed as the sum of atoms along its root-to-node path:

$$\mathbf{a}^{(i)} = \sum_{j \in \mathrm{path}(i)} \mathbf{d}^{(j)}, \tag{3}$$

where path(i) contains i and all its ancestors.

Morphological Description Generation

To supply complementary semantic cues for multi-modal fusion, we generate a fine-grained morphological description for each image using the open-source vision-language model Qwen3-VL-2B-Instruct[27]. To strictly prevent species-name or taxonomic label leakage, the VLM receives only the raw image (without any metadata or class labels) along with a constrained system prompt: *"Describe the visible morphological traits of the insect in this image in 2–3 concise sentences, focusing strictly on body segmentation, coloration, wing venation, antennal structure, and leg morphology. Do not guess or mention any species, genus, family, or common names."* We use deterministic greedy decoding. Furthermore, an automated post-processing regex filter checks all generated texts against the complete list of 162 Latin binomials, genera, families, orders, and common names, masking any accidental taxonomic terms with `[insect]`. An audit across the entire dataset confirmed 0.0% species-name leakage after filtering. The description $\mathbf{T}$ is then encoded by a frozen CLIP ViT-L/14 text encoder into a semantic embedding:

$$\mathbf{f}_{\mathrm{txt}} = \mathscr{E}_{\mathrm{CLIP}}(\mathbf{T}) \in \mathbb{R}^{D_t}. \tag{4}$$

For training images, descriptions and their CLIP embeddings are pre-computed offline once. For unseen test images during online evaluation, Qwen3-VL generates the morphological description on-the-fly prior to CLIP text encoding.

Online Hierarchical Sparse Pursuit and Multi-Modal Classification

In the online phase, GEPC processes each input image through a fixed sequence: hierarchical sparse pursuit, feature reconstruction, multi-modal fusion, and classification. All foundation encoders remain frozen.

Hierarchical Sparse Pursuit via Non-Negative HB-OMP. Given the DINOv2 visual feature $\mathbf{f}_{\mathrm{vis}} \in \mathbb{R}^{D_v}$ and the offline dictionary $\mathbf{D} \in \mathbb{R}^{D_v \times |\mathscr{N}|}$, we solve a path-constrained sparse coding problem using HB-OMP. Unlike standard OMP, which selects atoms globally, HB-OMP restricts each expansion step to the children of the current path endpoint. Formally, we seek a non-negative sparse code $\alpha \in \mathbb{R}_+^{|\mathscr{N}|}$ minimizing:

$$\min_{\alpha \geq \mathbf{0}} \|\mathbf{f}_{\mathrm{vis}} - \mathbf{D}\alpha\|_2^2 \quad \text{s.t.} \quad \mathrm{supp}(\alpha) \text{ forms a complete root-to-leaf path.} \tag{5}$$

To guarantee that every recovered support $\mathscr{S}$ corresponds to a complete three-level Order $\rightarrow$ Family $\rightarrow$ Species path (rather than terminating prematurely at an internal node), we execute Algorithm 1 for exactly $L = 3$ hierarchy levels rather than stopping early on an arbitrary residual tolerance. At each level $\ell \in \{1,2,3\}$, candidate child atoms are scored via directional cosine similarity $c_i = \max\left(0, \frac{\langle \mathbf{r}, \mathbf{d}^{(i)} \rangle}{\|\mathbf{r}\|_2 \|\mathbf{d}^{(i)}\|_2}\right)$ (with ties broken by the direct prototype similarity $\langle \mathbf{f}_{\mathrm{vis}}, \mathbf{a}^{(i)} \rangle$), and non-negative least squares (NNLS) updates the coefficients over the active path support.

Visual Feature Reconstruction. Rather than using α^* directly, we reconstruct the visual component strictly aligned with the chosen taxonomic path:

$$\mathbf{f}_{\mathrm{vis}}^{\mathrm{rec}} = \mathbf{D}\alpha^* \in \mathbb{R}^{D_v}. \tag{6}$$

This denoised, hierarchy-constrained representation filters out background clutter and taxonomically irrelevant variations, preparing a structured visual signal for multi-modal fusion.

Multi-Modal Fusion and Classification. The reconstructed visual feature is concatenated with the text embedding $\mathbf{f}_{\mathrm{txt}} \in \mathbb{R}^{D_t}$ to form a joint representation $\mathbf{f}_{\mathrm{fuse}} = \mathbf{f}_{\mathrm{vis}}^{\mathrm{rec}} \oplus \mathbf{f}_{\mathrm{txt}} \in \mathbb{R}^{D_v + D_t}$. A lightweight two-layer MLP projection head $\mathscr{P} : \mathbb{R}^{D_v + D_t} \rightarrow \mathbb{R}^{D_z}$ and a linear classifier $\mathscr{C} : \mathbb{R}^{D_z} \rightarrow \mathbb{R}^K$ produce the final species prediction:

$$\hat{\mathbf{y}} = \mathrm{softmax}\Big(\mathscr{C}\big(\mathscr{P}(\mathbf{f}_{\mathrm{fuse}})\big)\Big) \in \mathbb{R}^K, \tag{7}$$

where $K = 162$ is the number of pest species.

Training. Only $\mathscr{P}$ and $\mathscr{C}$ are trained, with all foundation encoders and the dictionary $\mathbf{D}$ frozen. We optimize the standard cross-entropy loss:

$$\mathscr{L}_{\mathrm{cls}} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} y_k^{(n)} \log \hat{y}_k^{(n)}, \tag{8}$$

where $y_k^{(n)} \in \{0,1\}$ is the ground-truth species indicator.

Algorithm 1 Hierarchical Beam OMP (HB-OMP) Inference with Non-Negative Leaf Guarantee

Require: Visual feature $\mathbf{f}_{\text{vis}}$, dictionary $\mathbf{D}$, node prototypes $\{\mathbf{a}^{(i)}\}$, children map $\text{chi}(\cdot)$, beam size B

Initialize hypothesis set $\mathscr{H} \leftarrow \{(\text{root}, \mathbf{f}_{\text{vis}}, \varnothing, \mathbf{0})\}$
for level $\ell = 1$ to 3 (Order, Family, Species) **do**
  $\mathscr{H}_{\text{new}} \leftarrow \varnothing$
  for each $h = (\text{node}, \mathbf{r}, \mathscr{S}, \alpha) \in \mathscr{H}$ **do**
    $\mathscr{I}_{\text{cand}} \leftarrow \text{chi}(\text{node})$
    if $|\mathscr{I}_{\text{cand}}| = 1$ and $\|\mathbf{d}^{(i)}\|_2 = 0$ for $i \in \mathscr{I}_{\text{cand}}$ **then**
      Add $(i, \mathbf{r}, \mathscr{S} \cup \{i\}, \alpha)$ to $\mathscr{H}_{\text{new}}$ {Deterministic single-child transition}
    else
      Compute directional score $c_i \leftarrow \max\left(0, \frac{\langle \mathbf{r}, \mathbf{d}^{(i)} \rangle}{\|\mathbf{r}\|_2 \|\mathbf{d}^{(i)}\|_2}\right)$ for $i \in \mathscr{I}_{\text{cand}}$
      Select top-B indices $\mathscr{I}_{\text{top}}$ from $\mathscr{I}_{\text{cand}}$ by descending c_i (breaking ties via $\langle \mathbf{f}_{\text{vis}}, \mathbf{a}^{(i)} \rangle$)
      for $i \in \mathscr{I}_{\text{top}}$ **do**
        $\mathscr{S}' \leftarrow \mathscr{S} \cup \{i\}$
        $\alpha' \leftarrow \arg\min_{\alpha \geq \mathbf{0}: \text{supp}(\alpha) \subseteq \mathscr{S}'_+} \|\mathbf{f}_{\text{vis}} - \mathbf{D}\alpha\|_2^2$, where $\mathscr{S}'_+ = \{j \in \mathscr{S}' : \|\mathbf{d}^{(j)}\|_2 > 0\}$
        $\mathbf{r}' \leftarrow \mathbf{f}_{\text{vis}} - \mathbf{D}\alpha'$
        Add $(i, \mathbf{r}', \mathscr{S}', \alpha')$ to $\mathscr{H}_{\text{new}}$
      end for
    end if
  end for
  $\mathscr{H} \leftarrow$ top B hypotheses in $\mathscr{H}_{\text{new}}$ with smallest $\|\mathbf{r}\|_2^2$
end for
return Sparse coefficient α^* and complete root-to-leaf path $\mathscr{S}^*$ from hypothesis with minimal residual

Experiments

Datasets, Protocol, and Implementation Details

Dataset and Split Protocol. We evaluate on the AP162 dataset[17], a large-scale agricultural pest benchmark comprising 162 species and 194,700 field images. Following the official benchmark protocol[17], the dataset is partitioned into training (60%, 116,822 images), validation (10%, 19,470 images), and test (30%, 58,408 images) splits using stratified sampling across all 162 classes. The dataset exhibits severe long-tailed imbalance: per-class sample counts range from 61 to 6,010. To directly evaluate long-tailed recognition performance, we partition the 162 categories based on their real training sample counts n_c into three standard subsets: **Head classes** ($n_c > 1{,}000$, 35 classes), **Medium classes** ($300 \leq n_c \leq 1{,}000$, 83 classes), and **Tail classes** ($n_c < 300$, 44 classes).

Implementation Details. All experiments are implemented in PyTorch and executed on two NVIDIA RTX A6000 (48 GB) GPUs. To account for statistical variability, we run all experiments across 3 independent random seeds ($\{42, 1024, 2025\}$) and report mean $\pm$ standard deviation. We adopt frozen DINOv2 ViT-L/14[13] and CLIP ViT-L/14[10] as visual and text backbones, producing $D_v = 1{,}024$ and $D_t = 768$ feature embeddings, respectively. The fusion projection head $\mathscr{P}$ is a two-layer MLP ($1792 \to 512 \to 256$) with LayerNorm, ReLU activation, and dropout ($p = 0.3$), followed by a linear classifier $\mathscr{C}$ ($256 \to 162$). We optimize $\mathscr{P}$ and $\mathscr{C}$ using AdamW (learning rate 5×10^{-4}, weight decay 10^{-4}) and cosine annealing over 100 epochs with a batch size of 256. For HB-OMP, we set beam size $B = 4$.

Formal Definition of Evaluation Metrics. To eliminate ambiguity regarding metric aggregation and path evaluation, we formally define all reported metrics over the N_{test} test samples and $K = 162$ classes:

- **Overall Accuracy (Acc):** Standard top-1 micro-averaged accuracy across all test samples, $\text{Acc} = \frac{1}{N_{\text{test}}} \sum_{n=1}^{N_{\text{test}}} \mathbb{I}(\hat{y}^{(n)} = y^{(n)})$.
- **Macro Precision (Prec), Macro Recall (Rec), and Macro F1:** Unweighted arithmetic means of per-class precision, recall, and F1-score across all $K = 162$ classes. We additionally report macro recall separately on the Head (Rec_{Head}), Medium (Rec_{Med}), and Tail (Rec_{Tail}) subsets.
- **Pursuit Path Accuracy ($\text{PA}_{\text{pursuit}}$) and Classifier Path Accuracy (PA_{cls}):** Let $\mathscr{T}(y) = (o^*, f^*, s^*)$ be the ground-truth taxonomic path of label y. $\text{PA}_{\text{pursuit}}$ measures the fraction of test samples where the visual HB-OMP support $\mathscr{S}^*$ exactly matches $\mathscr{T}(y)$ at all three levels. Conversely, PA_{cls} measures the hierarchical path accuracy induced by the final classifier prediction $\hat{y}$, defined as the average node-level correctness across Order, Family, and Species: $\text{PA}_{\text{cls}} = \frac{1}{3N_{\text{test}}} \sum_{n=1}^{N_{\text{test}}} \left[\mathbb{I}(o_{\text{cls}}^{(n)} = o^{*(n)}) + \mathbb{I}(f_{\text{cls}}^{(n)} = f^{*(n)}) + \mathbb{I}(s_{\text{cls}}^{(n)} = s^{*(n)})\right]$. Note that because the visual encoder, dictionary

D, and HB-OMP solver are frozen, $\mathrm{PA}_{\mathrm{pursuit}}$ depends solely on the pursuit configuration (88.9% at $B = 4$), whereas $\mathrm{PA}_{\mathrm{cls}}$ reflects downstream fusion improvements (91.8% for full GEPC). In our ablation tables, we report both $\mathrm{PA}_{\mathrm{pursuit}}$ and $\mathrm{PA}_{\mathrm{cls}}$ to clarify why downstream fusion affects classifier-level hierarchy accuracy while leaving visual pursuit paths unchanged.

- **Path–Classifier Agreement (PC):** The proportion of test samples for which the leaf species s_{hb} selected by HB-OMP matches the final species $\hat{y}$ predicted by the multimodal classifier: $\mathrm{PC} = \frac{1}{N_{\mathrm{test}}} \sum_{n=1}^{N_{\mathrm{test}}} \mathbb{I}(s_{\mathrm{hb}}^{(n)} = \hat{y}^{(n)})$.

Main Results

Table 1. Performance comparison of different methods on the AP162 test set (mean ± SD over 3 random seeds under matched splits). Best results are in **bold**.

Method	Training Data	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	$\mathrm{Rec}_{\mathrm{Head}}$ (%)	$\mathrm{Rec}_{\mathrm{Med}}$ (%)	$\mathrm{Rec}_{\mathrm{Tail}}$ (%)
MobileNetV3-large[28]	Original	75.8 ± 0.5	77.9 ± 0.6	71.9 ± 0.6	73.2 ± 0.5	86.4	73.1	54.2
ResNet-18[29]	Original	77.5 ± 0.4	78.9 ± 0.5	74.0 ± 0.5	75.1 ± 0.4	87.8	75.4	56.8
ViT-B/16[30]	Original	77.6 ± 0.4	79.4 ± 0.5	74.2 ± 0.5	75.3 ± 0.5	88.1	75.6	57.0
ViT-L/16[30]	Original	80.7 ± 0.4	81.2 ± 0.4	77.3 ± 0.5	77.5 ± 0.4	90.2	78.9	60.9
ResNet-50[29]	Original	82.5 ± 0.3	81.8 ± 0.4	79.3 ± 0.4	80.2 ± 0.4	91.0	81.1	63.9
ConvMixer-20[31]	Original	85.2 ± 0.4	85.2 ± 0.4	82.5 ± 0.5	83.1 ± 0.4	92.5	84.3	69.0
DenseNet-161[32]	Original	85.8 ± 0.3	86.0 ± 0.4	83.3 ± 0.4	84.2 ± 0.3	92.9	85.1	70.3
SwinTransformer-base[33]	Original	88.7 ± 0.3	89.2 ± 0.3	85.9 ± 0.4	87.5 ± 0.3	94.2	87.8	74.2
AP162 (Wang et al.)[17]	Original	89.2 ± 0.3	89.4 ± 0.3	86.9 ± 0.4	87.7 ± 0.3	94.6	88.7	75.9
CLIP-ViT-L/14 (zero-shot)[10]	–	6.48	3.76	4.15	2.93	7.1	4.0	1.1
Qwen3-VL-2B (zero-shot)[27]	–	3.55	–	–	–	–	–	–
DINOv2-ViT-L/14 + MLP (matched head)	Original	85.4 ± 0.3	85.9 ± 0.4	82.9 ± 0.4	84.4 ± 0.3	93.4	85.0	68.5
DINOv2-ViT-L/14 + MLP + CB-Focal[21]	Original	86.3 ± 0.3	86.5 ± 0.4	84.6 ± 0.4	85.5 ± 0.4	92.8	86.1	73.5
DINOv2-ViT-L/14 + MLP (matched head)	Augmented	87.9 ± 0.3	88.3 ± 0.3	85.9 ± 0.4	87.1 ± 0.3	94.0	87.6	76.7
HCEP (DINOv2-ViT-L/14)[12]	Original	86.2 ± 0.4	86.7 ± 0.4	83.6 ± 0.5	85.1 ± 0.4	93.6	85.7	69.7
HCEP (DINOv2-ViT-L/14)[12]	Augmented	88.4 ± 0.3	88.7 ± 0.3	86.4 ± 0.4	87.5 ± 0.3	94.3	88.0	77.5
GEPC (Ours, w/o synthetic data)	Original	87.4 ± 0.3	87.8 ± 0.3	84.5 ± 0.4	86.1 ± 0.3	94.1	86.8	70.8
GEPC (Ours, full)	Augmented	**90.4 ± 0.3**	**90.1 ± 0.3**	**88.3 ± 0.4**	**89.2 ± 0.3**	**95.1**	**89.6**	**82.4**

Comparison with State-of-the-Art and Matched-Capacity Baselines. As shown in Table 1, all supervised baselines were trained and evaluated under the exact same 6:1:3 data split and evaluation protocol. CNN and ViT-based models achieve competitive overall accuracy, yet their macro recall—especially on the 44 tail classes ($\mathrm{Rec}_{\mathrm{Tail}}$)—is severely degraded by class imbalance. The multi-model ensemble in AP162 (Wang et al.)[17] reaches 89.2% overall accuracy and 75.9% tail recall. In zero-shot evaluations, CLIP-ViT-L/14 (6.48%) and Qwen3-VL-2B (3.55%) perform poorly, confirming that general-domain vision-language pre-training does not directly transfer to fine-grained entomological taxonomy without specialized adaptation.

To rigorously isolate the architectural contributions of GEPC from backbone capacity and training data volume, Table 1 also evaluates matched-capacity baselines using the identical frozen DINOv2-ViT-L/14 backbone:

1. **Matched DINOv2 + MLP and Standard Class-Balancing:** Replacing the linear probe (84.5%, Table 2) with the exact same two-layer MLP head used in GEPC yields 85.4% accuracy on the original split. Pairing DINOv2+MLP with Class-Balanced Focal Loss (CB-Focal)[21] raises tail recall from 68.5% to 73.5%, whereas augmenting DINOv2+MLP with our filtered synthetic dataset boosts tail recall to 76.7% and overall accuracy to 87.9%.

2. **Controlled Comparison with HCEP[12]:** Under the same DINOv2 backbone and original training split, our GEPC without synthetic augmentation (87.4% Acc, 84.5% Rec) outperforms HCEP (86.2% Acc, 83.6% Rec) by +1.2% accuracy. When both methods are trained with identical synthetic augmentation, GEPC (90.4% Acc, 88.3% Rec, 82.4% $\mathrm{Rec}_{\mathrm{Tail}}$) surpasses augmented HCEP (88.4% Acc, 86.4% Rec, 77.5% $\mathrm{Rec}_{\mathrm{Tail}}$) by +2.0% overall accuracy and +4.9% tail-class recall. This confirms that our non-negative directional pursuit and morphological text fusion provide substantial gains beyond the inherited HCEP dictionary structure and data augmentation alone.

Qualitative Analysis and Graceful Taxonomic Degradation. Fig. 3 presents qualitative visualizations on three representative pest species. The top panel (a) displays synthetic images generated by our semantics-guided diffusion pipeline, demonstrating ecological plausibility and morphological fidelity. The bottom panel (b) presents the top-3 candidate paths recovered by HB-OMP on real test images, where bar lengths denote normalized path weights. Quantitatively, across the test set the visual pursuit leaf agrees with the final prediction on 94.6% of samples; even when they disagree, the pursuit path remains correct at the family (97.8%) and order (99.1%) levels, so the visual stream degrades gracefully to coarse-grained taxonomic predictions rather than failing outright.

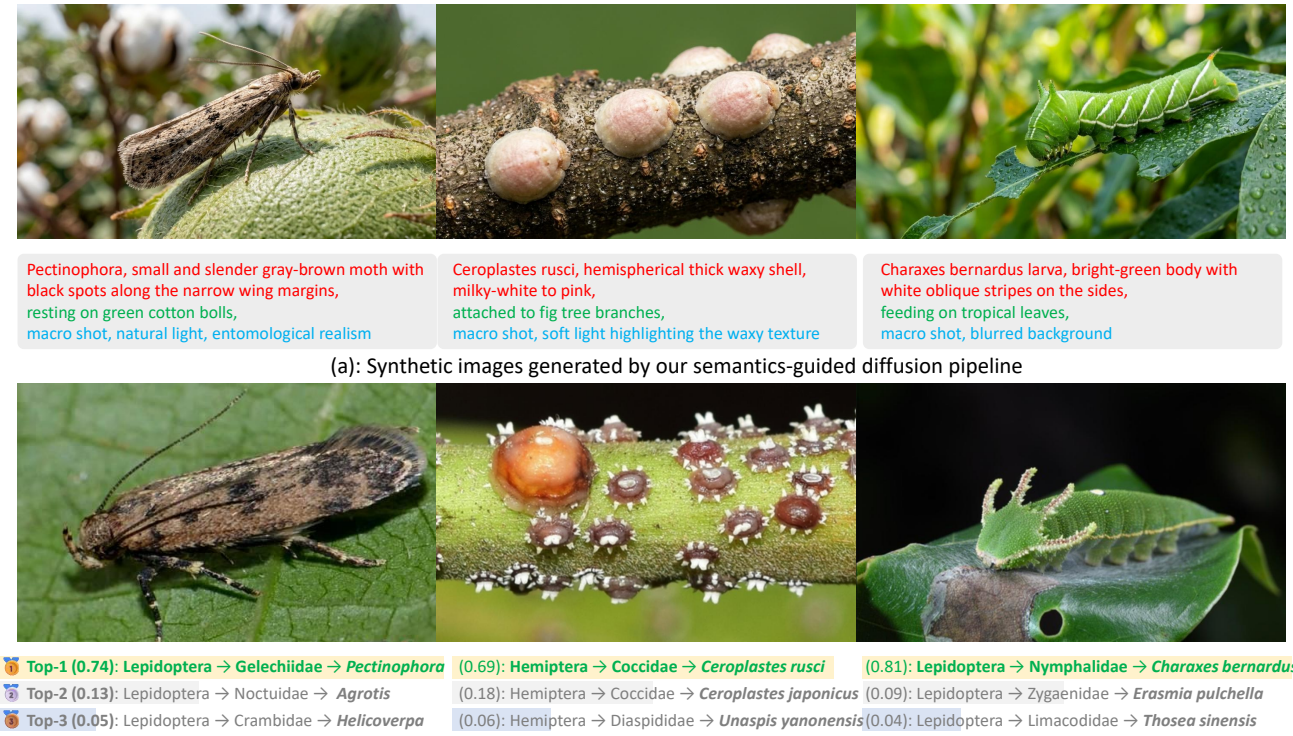


Figure 3. Qualitative analysis of GEPC on three representative pest species. (a) Synthetic images generated by our semantics-guided diffusion pipeline. (b) Top-3 hierarchical inference paths decoded from HB-OMP on real test images.

Table 2. Incremental ablation of GEPC components on AP162 (all configurations use the exact same 6:1:3 split and hyperparameters).

Configuration	Acc (%)	Prec (%)	Rec (%)	F1 (%)	$PA_{pursuit}$ (%)	PA_{cls} (%)	PC (%)
(1) Base (frozen DINOv2 + linear, original data)	84.5	85.0	82.0	83.5	–	88.2	–
(2) + Synthetic data augmentation	87.5	88.0	85.5	86.7	–	90.1	–
(3) + Dual-modal fusion (MLP head)	88.8	89.2	87.0	88.1	–	90.9	–
(4) + **HB-OMP (full GEPC)**	**90.4**	**90.1**	**88.3**	**89.2**	**88.9**	**91.8**	**94.6**

Ablation Studies

Diagnosing the Core Challenges. We empirically validate the two bottlenecks motivating our design: long-tailed data distribution and high inter-class visual similarity. As shown in Fig. 4, which plots per-class test precision against test sample counts (proportional to training sample sizes under the stratified 6:1:3 split) across representative classes, recognition precision correlates strongly with sample availability; rare classes lag far behind the dataset average. However, data scarcity is not the sole bottleneck—visual ambiguity persists even for well-sampled categories. For instance, among three morphologically similar *Spodoptera* species with comparable test-set sizes (637, 625, and 779 images), *S. frugiperda* achieves 83.36% precision on a fine-tuned end-to-end Swin Transformer baseline, while *S. litura* and *S. exigua* reach 94.79% and 90.77%, respectively—confirming that fine-grained visual distinctions remain a major challenge.

Incremental Component Ablation and Baseline Reconciliation. To quantify the contribution of each component, we incrementally activate them over a frozen DINOv2 backbone with a linear classifier in Table 2. We also clarify the relationship between Table 2 and Table 3: Configuration (1) in Table 2 (84.5% Acc) is the bare visual baseline (*frozen DINOv2 + linear classifier, without dual-modal fusion or HB-OMP*) trained on original data, whereas the "Original only" row in Table 3 (87.4% Acc) evaluates the *complete GEPC architecture (including HB-OMP pursuit, morphological text fusion, and the MLP head)* trained exclusively on original data without synthetic augmentation. Introducing synthetic data augmentation in Configuration (2) brings the largest recall gain (+3.5%), directly mitigating the long-tail bottleneck. Dual-modal fusion in Configuration (3) further improves recall by +1.5% as CLIP-encoded morphological descriptions disambiguate visually similar species. Finally, equipping the model with HB-OMP and the differential dictionary in Configuration (4) adds +1.3% recall, achieves 88.9%

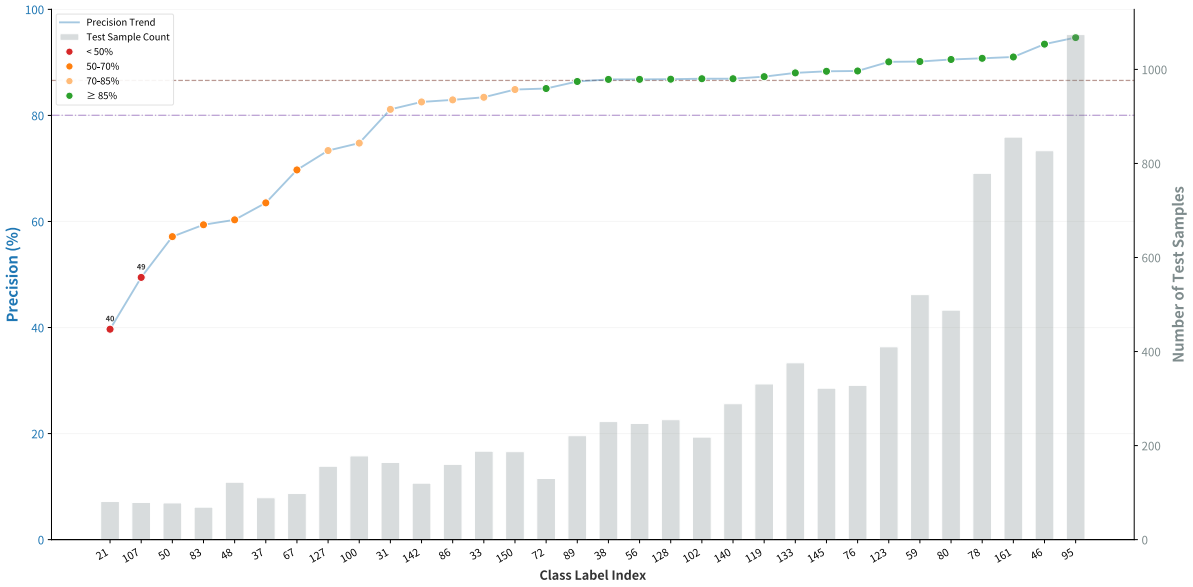


Figure 4. Per-class test precision versus test sample count across representative AP162 categories.

exact visual pursuit path accuracy ($PA_{pursuit}$), 91.8% classifier hierarchy accuracy (PA_{cls}), and 94.6% path–classifier agreement (PC).

Further Analysis

Effect of Synthetic Data Quality Control. To isolate the effect of automated CLIP filtering from manual screening, we compare four training data configurations under the full GEPC architecture in Table 3: (a) original training data only (116,822 images), (b) adding all 96,400 raw synthetic images without filtering, (c) adding the 78,120 synthetic images retained by automated CLIP filtering ($\tau_{CLIP} = 0.30$) alone, and (d) our full pipeline combining CLIP filtering with manual artifact screening (71,650 retained synthetic images). While naive augmentation improves macro recall by +1.3%, automated CLIP filtering alone yields the majority of the gain (+3.3% recall, reaching 89.9% Acc), and subsequent manual screening provides a further +0.5% recall refinement (90.4% Acc) by eliminating subtle anatomical distortions.

Table 3. Effect of synthetic data quality control on the full GEPC architecture.

Training Data	Added Synthetic Images	Acc (%)	Prec (%)	Rec (%)	F1 (%)
Original only	0	87.4	87.8	84.5	86.1
+ Naive synthetic (unfiltered)	96,400	88.0	88.3	85.8	87.0
+ CLIP-filtered only ($\tau_{CLIP} = 0.30$)	78,120	89.9	89.7	87.8	88.7
+ CLIP-filtered + manual screening (Full)	71,650	**90.4**	**90.1**	**88.3**	**89.2**

Dictionary Type and Non-Negative Pursuit Constraints. We contrast our differential dictionary with an absolute embedding dictionary under the same HB-OMP pursuit, and also evaluate the effect of removing the directional non-negativity constraint ($\alpha \geq \mathbf{0}$) in Table 4. Compared to absolute embeddings (78.4% PA_{cls}), the differential formulation explicitly encodes parent–child visual deviations. Moreover, enforcing $\alpha \geq \mathbf{0}$ increases exact visual pursuit accuracy $PA_{pursuit}$ from 84.2% to 88.9% and path–classifier agreement PC from 89.5% to 94.6% by eliminating sign-flipping sibling ambiguities in two-species families.

Beam Size Sensitivity. We evaluate the impact of beam size B in HB-OMP (Table 5). Both classification accuracy and path accuracy improve from $B = 1$ to $B = 4$, and slightly degrade or saturate thereafter as wider beams occasionally admit visually noisy outlier branches. We therefore adopt $B = 4$ as the default setting.

Multi-Modal Fusion Strategy and Runtime Cost. We compare two fusion mechanisms for combining the hierarchy-reconstructed visual feature $\mathbf{f}_{vis}^{rec}$ and the text embedding $\mathbf{f}_{txt}$: gated fusion and direct concatenation followed by the two-layer MLP $\mathscr{P}$. Table 6 shows that concatenation consistently outperforms gated fusion. Importantly, Table 6 also clarifies why PA_{cls} increases from 90.4% to 91.8% while $PA_{pursuit}$ remains constant at 88.9%: because DINOv2, $\mathbf{D}$, and HB-OMP precede the

Table 4. Effect of dictionary construction and non-negative pursuit constraints.

Dictionary & Pursuit Variant	Acc (%)	Prec (%)	Rec (%)	F1 (%)	$\mathrm{PA}_{\mathrm{pursuit}}$ (%)	$\mathrm{PA}_{\mathrm{cls}}$ (%)	PC (%)
Absolute embeddings	86.5	86.9	83.8	85.3	74.1	78.4	83.2
Differential (unconstrained LS, $\lvert\langle \mathbf{r}, \mathbf{d} \rangle\rvert$)	89.3	89.4	87.2	88.3	84.2	90.5	89.5
Differential (directional + NNLS, Ours)	**90.4**	**90.1**	**88.3**	**89.2**	**88.9**	**91.8**	**94.6**

Table 5. Effect of beam size B in HB-OMP.

Beam size B	Acc (%)	Prec (%)	Rec (%)	F1 (%)	$\mathrm{PA}_{\mathrm{pursuit}}$ (%)	$\mathrm{PA}_{\mathrm{cls}}$ (%)	PC (%)
1	87.8	88.5	85.0	86.7	85.4	88.6	92.1
2	88.5	89.0	86.5	87.7	87.1	90.2	93.4
4	**90.4**	**90.1**	**88.3**	**89.2**	**88.9**	**91.8**	**94.6**
6	89.1	90.0	88.2	89.1	88.7	91.7	94.4
8	89.1	90.0	88.1	89.0	88.6	91.6	94.3

fusion head, the visual pursuit path $\mathscr{S}^*$ is identical under both fusion strategies ($\mathrm{PA}_{\mathrm{pursuit}} = 88.9\%$), whereas concatenation enables the downstream classifier $\mathscr{C}$ to recover more accurate species and family predictions ($\mathrm{PA}_{\mathrm{cls}} = 91.8\%$) and achieve higher agreement (PC $= 94.6\%$) with the visual pursuit leaf.

In terms of per-image online inference latency on a single NVIDIA RTX A6000 GPU (batch size 1), DINOv2 ViT-L/14 visual feature extraction takes 9.4 ms, HB-OMP pursuit ($B = 4$) takes 3.8 ms, and MLP fusion plus classification takes 0.4 ms (totaling 13.6 ms when using visual features or cached descriptions). On-the-fly morphological caption generation via Qwen3-VL-2B-Instruct requires 142.0 ms plus 3.1 ms for CLIP text encoding, bringing end-to-end multimodal inference to 158.7 ms per image (~ 6.3 FPS), which is well suited for field pest diagnostic applications.

Table 6. Comparison of multi-modal fusion strategies (distinguishing upstream visual pursuit accuracy $\mathrm{PA}_{\mathrm{pursuit}}$ from downstream classifier hierarchy accuracy $\mathrm{PA}_{\mathrm{cls}}$).

Fusion Method	Acc (%)	Prec (%)	Rec (%)	F1 (%)	$\mathrm{PA}_{\mathrm{pursuit}}$ (%)	$\mathrm{PA}_{\mathrm{cls}}$ (%)	PC (%)
Gated fusion	88.5	89.0	86.8	87.9	88.9	90.4	93.1
Concatenation	**90.4**	**90.1**	**88.3**	**89.2**	**88.9**	**91.8**	**94.6**

Conclusion

In this paper, we proposed GEPC, a unified framework for agricultural pest recognition that simultaneously addresses class imbalance, fine-grained visual ambiguity, and hierarchical interpretability through three complementary modules: a semantics-guided diffusion pipeline with structured prompting and CLIP-based quality filtering for generating balanced, high-fidelity synthetic training samples; a hierarchical concept pursuit module that reformulates visual feature extraction as a non-negative path-constrained sparse recovery problem over a verified Order–Family–Species differential dictionary; and a dual-modal alignment module that fuses hierarchy-reconstructed DINOv2 visual features with species-anonymized CLIP morphological embeddings. Extensive experiments on the AP162 benchmark demonstrate that GEPC achieves 90.4% overall accuracy, 82.4% tail-class recall, and 94.6% agreement between the visual pursuit leaf and the final multimodal prediction, offering a scalable and transparent solution for intelligent pest management.

References

1. Market.us. Insect pest control market (2026).
2. Du, M. *et al.* Improving long-tailed pest classification using diffusion model-based data augmentation. *Comput. Electron. Agric.* **234**, 110244 (2025).
3. Agrawal, N., Mahrishi, M., Gupta, M. K. & Bohra, M. K. Large scale multi-class pest image classification using structurally adapted densenet architecture. *Sci. Reports* **16**, 20011 (2026).

4. Chen, S., Gao, Q. & He, Y. A pest image recognition method for long-tail distribution problem. *Front. Environ. Sci.* **12** (2024).
5. Li, X. *et al.* Pestvl-net: Enabling multimodal pest learning via fine-grained vision-language interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8826–8835 (2026).
6. Zhang, W., Fang, T., Lu, L., Wang, L. & Zhong, W. Cpj: Explainable agricultural pest diagnosis via caption–prompt–judge with llm-judged refinement. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 5156–5160 (2026).
7. Liu, J. & Wang, X. Plant diseases and pests detection based on deep learning: a review. *Plant Methods* **17**, 22 (2021).
8. Shafik, W., Tufail, A., De Silva, L. C. & Apong, R. A. A. H. M. Lightweight vision transformer and resnet-9 models for real-time plant disease detection and pest classification with shap explainability. *BMC Plant Biol.* **26**, 838 (2026).
9. Sharma, N., Al-Yarimi, F. A. M., Bharany, S., Rehman, A. U. & Taye, B. M. Hybrid ai model with cnns and vision transformers for precision pest classification in crops. *Food Sci. & Nutr.* **13**, e71174 (2025).
10. Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763 (2021).
11. Nawaz, U. *et al.* Agriclip: Adapting clip for agriculture and livestock via domain-specialized cross-model alignment. In *Proceedings of the International Conference on Computational Linguistics*, 9630–9639 (2025).
12. Nguyen, N., Ding, T. & Vidal, R. Hierarchical concept embedding & pursuit for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2907–2917 (2026).
13. Oquab, M. *et al.* DINOv2: Learning Robust Visual Features Without Supervision. *Transactions on Mach. Learn. Res.* (2024).
14. Xie, C. *et al.* Automatic classification for field crop insects via multiple-task sparse representation and multiple-kernel learning. *Comput. Electron. Agric.* **119**, 123–132 (2015).
15. Venugoban, K. & Ramanan, A. Image classification of paddy field insect pests using gradient-based features. *Int. J. Mach. Learn. Comput.* **4**, 1–5 (2014).
16. Wu, X., Zhan, C., Lai, Y.-K., Cheng, M.-M. & Yang, J. Ip102: A large-scale benchmark dataset for insect pest recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8779–8788 (2019).
17. Wang, K. *et al.* Ap162: A large-scale dataset for agricultural pest recognition. *Comput. Electron. Agric.* **237**, 110520 (2025).
18. De Cesaro Júnior, T., de Oliveira, C. A. L., Lau, D. & Rieder, R. Fine-grained recognition of insect pests from digital images: A survey. *Neotropical Entomol.* **55**, 48 (2026).
19. Lu, H. *et al.* A survey on deep learning-based object detection for crop monitoring: pest, yield, weed, and growth applications. *The Vis. Comput.* **41**, 10069–10094 (2025).
20. Ahsan, F. F., Thomas, M. L., Laga, H. & Sohel, F. A semi-supervised approach for classifying insect developmental phases from repurposed ip102. *Comput. Electron. Agric.* **242**, 111337 (2026).
21. Cui, Y., Jia, M., Lin, T.-Y., Song, Y. & Belongie, S. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9268–9277 (2019).
22. Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, 618–626 (2017).
23. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the International Conference on Neural Information Processing Systems*, 4768–4777 (2017).
24. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695 (2022).
25. OpenAI. Chatgpt (gpt-4) [large language model]. https://chat.openai.com/chat (2023). Accessed: 2024-05-20.
26. EPPO. EPPO Global Database. *Eur. Mediterr. Plant Prot. Organ.* (2025).
27. Bai, S. *et al.* Qwen3-vl technical report (2025). 2511.21631.
28. Howard, A. *et al.* Searching for mobilenetv3. In *IEEE/CVF International Conference on Computer Vision*, 1314–1324 (2019).

29. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).

30. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* (2021).

31. Trockman, A. & Kolter, J. Z. Patches are all you need? *Transactions on Mach. Learn. Res.* (2023).

32. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2261–2269 (2017).

33. Liu, Z. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision*, 9992–10002 (2021).

Author contributions statement

Y.X. and Y.S. conceived the study and designed the methodology. Y.X. implemented the codebase, conducted the data generation and classification experiments, and prepared the figures. Y.S. supervised the project, verified the taxonomic annotations, and led the manuscript revision. Both authors reviewed and approved the final manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability statement

The AP162 benchmark dataset analyzed during the current study is publicly available as described by Wang et al.[17].

Code availability

The implementation of GEPC will be provided by the corresponding author upon reasonable request.

Additional information

Competing interests

The authors declare no competing interests.