REFERENCE-CALIBRATED EXPERT COMPOSITION FOR OPEN-DOMAIN FINE-GRAINED RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

A long-standing goal of fine-grained visual recognition is to build a general-purpose system that recognizes fine-grained categories across a broad and growing range of domains. However, the visual cues needed to distinguish closely related categories vary across domains, and models specialized through task- or domain-specific training may perform less reliably outside their target domains. Extending coverage through this approach therefore requires repeated supervision and model adaptation, making broad fine-grained recognition difficult to scale. Visual models pretrained on large-scale data offer an alternative foundation, as their representations already capture rich fine-grained distinctions. Our analysis shows that the examined models have uneven capabilities across domains but exhibit complementary strengths. These observations motivate composing pretrained experts whose strengths differ across domains. We propose a training-free framework that reads domain-specific capability from a small set of labeled reference images and composes heterogeneous frozen experts accordingly. Specifically, references calibrate domain-specific expert contributions for visual comparison and image–name matching, while within-class feature statistics define the visual comparison geometry. Held-out reference predictions further calibrate the balance between visual and semantic evidence. All pretrained models remain frozen, allowing new domains and new experts to be incorporated without retraining. We evaluate the framework on 598 mixed-domain classes, 10,000 species, and 4,431 product identities, demonstrating the utility of reference-calibrated composition across heterogeneous recognition tasks.

1 INTRODUCTION

Fine-grained visual recognition distinguishes closely related categories, such as bird species, vehicle variants, and product identities, supporting species identification, product retrieval, and precise object interaction. Dedicated methods have made substantial progress within fixed domains (Lin et al., 2015; Yang et al., 2018; He et al., 2022), but extending this success to *open-domain fine-grained recognition* remains challenging. Here, queries have unknown coarse domains and are classified against a shared, reference-defined vocabulary spanning heterogeneous semantic domains. Such coverage could support embodied agents and robots that must recognize specific objects across diverse environments.

The challenge extends beyond vocabulary size. Discriminative cues vary across domains: bird species may differ in plumage and beak shape, whereas vehicle and aircraft variants require different structural details. Features that separate classes in one domain may be less informative in another or capture within-class variation elsewhere. Expanding coverage through task-specific training requires repeated supervision and adaptation, which can distort transferable features and reduce out-of-distribution robustness (Kumar et al., 2022; Wortsman et al., 2022).

Large-scale pretraining offers an alternative foundation. CLIP and DINOv2 provide transferable visual representations (Radford et al., 2021; Oquab et al., 2023), while BioCLIP2 demonstrates strong species-level discrimination (Gu et al., 2025). However, broad semantic coverage does not guarantee uniformly strong fine-grained recognition. Large vision–language models also face limitations in visual discrimination, vision–language alignment, semantic grounding, and knowledge coverage (Yu et al., 2026), while targeted fine-tuning can reduce previously supported capabilities (Zhu et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

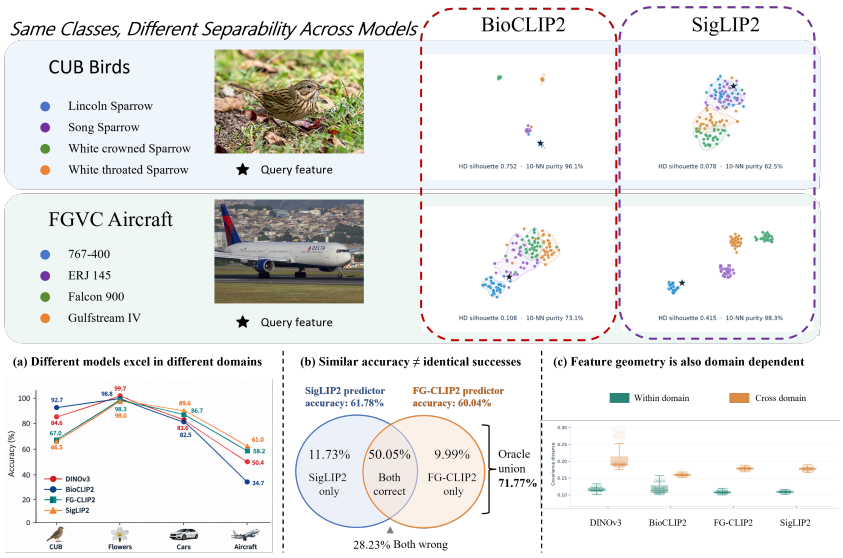


Figure 1: **Complementary expertise and domain-dependent feature geometry. Top:** Class separation varies across models, as reflected in silhouette scores (Rousseeuw, 1987) and neighborhood purity. **Bottom (a):** Frozen encoders exhibit different domain strengths. **Bottom (b):** SigLIP2 and FG-CLIP2 correctly classify partially different Aircraft queries(Avg. over multiple runs). **Bottom (c):** Within-class covariance estimates are more similar within domains than across domains.

2024). This raises a question: *how can we broaden fine-grained recognition by exploiting expertise already present in pretrained models?*

Our observations suggest that this expertise is unevenly distributed but complementary. Figure 1 (top) shows contrasting feature separation for neighboring bird and aircraft categories across pretrained models. The recognition results in bottom (a) agree: among the evaluated models, BioCLIP2 leads on CUB, DINOv3 on Flowers, and SigLIP2 on Cars and Aircraft. Complementarity also exists within a domain: SigLIP2 and FG-CLIP2 correctly classify partially different Aircraft queries in bottom (b). Together, these findings motivate *composing complementary pretrained capabilities* rather than relying on a single representation.

Domain dependence also affects how features should be compared. Within each encoder, trace-normalized within-class covariances from disjoint class groups are closer within domains than across domains (Figure 1, bottom (c)). This suggests domain-specific regularities in feature variation, motivating comparison geometry derived from pooled, class-centered reference residuals.

We propose *Reference-Calibrated Expert Composition*, a training-free framework that uses a few labeled references per class to determine *how much each expert contributes* and *how its features are compared*. *Reference-Derived Expert Competence* estimates domain-specific discriminability to assign soft expert weights. *Domain-Conditioned Comparison Geometry* uses regularized within-class statistics to compare queries with visual class prototypes. When class names are available, cross-fitted calibration of the branch weight balances visual-reference and class-name semantic evidence. At inference, retrieval supplies the coarse-domain context used to select these statistics and score the shared candidate vocabulary. New domains or experts can be incorporated by updating the reference library or expert pool and recalibrating the affected statistics, while all pretrained parameters remain frozen.

Our contributions are threefold:

- We study open-domain fine-grained recognition through *reference-calibrated expert composition*, exploiting complementary pretrained expertise and domain-dependent feature geometry.
- We introduce a training-free framework that derives soft expert weights, comparison geometry, and visual–semantic calibration from a few labeled references per class.

- We demonstrate its effectiveness across 598 mixed-domain classes, 10,000 species, and 4,431 product identities.

2 Related Work

Fine-grained recognition and pretrained expertise. Traditional approaches learn discriminative feature interactions and object parts within a target domain (Lin et al., 2015; Yang et al., 2018; He et al., 2022). More recent pretraining efforts strengthen biological recognition with hierarchical supervision in BioCLIP2 (Gu et al., 2025) and fine-grained vision–language alignment in FG-CLIP2 (Xie et al., 2026). IDEA (Ye et al., 2026) and UniFGVC (Guo et al., 2025) complement visual features with generated descriptions for training-free recognition. Our focus is on composing the domain-dependent capabilities already available in heterogeneous frozen representations.

Training-free adaptation. Tip-Adapter (Zhang et al., 2022) retrieves labeled support features from a key–value cache, APE (Zhu et al., 2023) refines feature priors, and TIP-X, introduced within SuS-X (Udandarao et al., 2023), improves support affinities through cross-modal information. GDA (Wang et al., 2024) combines a Gaussian discriminant classifier with zero-shot predictions, while ECALP (Li et al., 2025) exploits context-aware label propagation. These methods establish effective alternatives to parameter adaptation. Our focus is the reference calibration of heterogeneous experts for a shared candidate vocabulary, with domain-conditioned comparison geometry and separate visual and semantic expert weights.

Expert selection and composition. LogME (You et al., 2021) and PARC (Bolya et al., 2021) estimate transferability for model selection, whereas KNORA-E/U (Ko et al., 2008) select ensembles using local classifier competence. ZS-En and TF-En (Lu et al., 2024) customize vision–language ensembles for zero- and few-shot transfer. COSMIC (Huang et al., 2025) integrates CLIP and DINO representations through semantic graphs, and SOTA (Hu et al., 2026) balances multiple foundation models through self-adaptive optimal transport. Our framework uses a small labeled reference library to calibrate *both* which experts contribute and how their features are compared, with separate visual and semantic competence estimates and a cross-fitted mixture of the two evidence sources.

3 Method

Figure 2 summarizes the proposed framework. Given a small labeled reference library and a pool of frozen pretrained experts, our method operates in two stages. In *Stage I: Preparation*, the references are used to estimate domain-specific expert competence, construct a visual comparison geometry, and calibrate the relative contributions of visual and semantic evidence. These quantities are estimated offline and cached for a fixed reference library and expert pool. In *Stage II: Inference*, retrieval identifies a coarse-domain context for each query, and the corresponding cached statistics guide expert composition and fine-grained recognition.

3.1 Problem Setting

Let $\mathcal{C} = \{1, \dots, C\}$ denote the set of fine-grained candidate classes and $\mathcal{G}$ the set of coarse domains. Each class c has a known domain assignment $g(c) \in \mathcal{G}$, and $\mathcal{C}_g = \{c \in \mathcal{C} : g(c) = g\}$. Each class contains $K = 5$ labeled reference images $x_{c1}, \dots, x_{cK}$, forming the reference library $\mathcal{S}_{\text{ref}} = \{(x_{cj}, c) : c \in \mathcal{C},\ j = 1, \dots, K\}$. The class-to-domain mapping is known, whereas the domain of a new query is not given at test time.

The expert pool contains $M = 4$ frozen image encoders: DINOv3 (Siméoni et al., 2025), BioCLIP2 (Gu et al., 2025), FG-CLIP2 (Xie et al., 2026), and SigLIP2 (Tschannen et al., 2025). Expert m maps image x to the normalized feature $u_m(x) = f_m(x)/\|f_m(x)\|_2 \in \mathbb{R}^{D_m}$. BioCLIP2, FG-CLIP2, and SigLIP2 additionally provide frozen text encoders; their index set is denoted by $\mathcal{M}_T$. For class c, n_c denotes its class name. Calibration uses analytical reference statistics and discrete held-out selection of the branch weight, without gradient-based parameter updates. All pretrained encoders remain frozen.

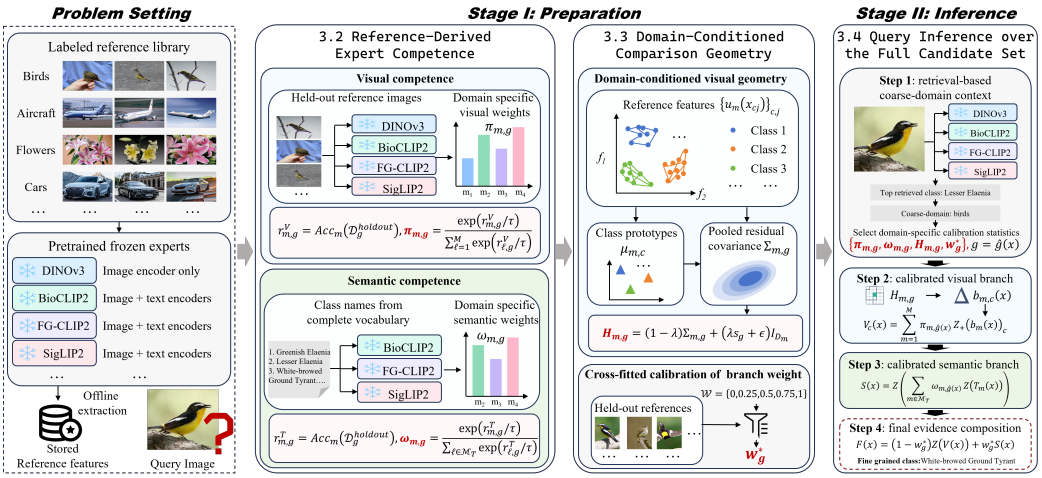


Figure 2: **Overview of the proposed framework.** Blue-shaded regions denote the visual branch, while green-shaded regions denote the semantic branch. In Stage I, reference replay estimates domain-specific visual and semantic expert competence, class-centered image features define a domain-conditioned comparison geometry, and cross-fitted reference predictions calibrate the visual–semantic branch weight. In Stage II, retrieval provides the coarse-domain context used to access the cached statistics. The visual and semantic branches then produce complementary scores that are combined for final prediction.

3.2 Reference-Derived Expert Competence

Different pretrained representations preserve different fine-grained distinctions. We therefore use the references to estimate, for each coarse domain, how reliably each frozen expert distinguishes categories through visual comparison and image–name matching.

Visual competence. We evaluate each image expert by predicting the class label of a held-out reference using the remaining references. For reference position j, the j-th image of every class is held out. When x_{cj} is treated as the query, expert m scores candidate class c' using the remaining $K-1$ references of that class:

$$a_{m,c'}^{(-j)}(x_{cj}) = \mathrm{MeanTop}_3 \left\{ u_m(x_{cj})^\top u_m(x_{c'k}) : k \neq j \right\}, \tag{1}$$

where $\mathrm{MeanTop}_3$ averages the three largest similarities. The visual competence of expert m in domain g is its average held-out top-1 accuracy within that domain:

$$r_{m,g}^V = \frac{1}{K|\mathcal{C}_g|} \sum_{c \in \mathcal{C}_g} \sum_{j=1}^{K} \mathbb{I}\left[c = \arg\max_{c' \in \mathcal{C}_g} a_{m,c'}^{(-j)}(x_{cj}) \right]. \tag{2}$$

This held-out cosine-retrieval accuracy provides a representation-level competence estimate before geometric calibration. We convert these estimates into soft domain-specific visual weights:

$$\pi_{m,g} = \frac{\exp(r_{m,g}^V/\tau)}{\sum_{\ell=1}^{M} \exp(r_{\ell,g}^V/\tau)}, \qquad \tau = 0.1. \tag{3}$$

Experts with higher held-out accuracy receive larger weights, and all image experts contribute to the visual branch through this soft mixture.

Semantic competence. Visual retrieval and image–name matching need not favor the same experts, so we calibrate the text-capable experts separately. Following the prompt-ensembling approach of CLIP (Radford et al., 2021), for each $m \in \mathcal{M}_T$, we average the text embeddings of seven generic photographic templates containing n_c and normalize the result to obtain the class-name prototype $t_{m,c}$. The image–name score is $T_{m,c}(x) = 100\, u_m(x)^\top t_{m,c}$. Matching the labeled reference images

to these prototypes gives the domain-specific semantic accuracy:

$$r^T_{m,g} = \frac{1}{K|\mathcal{C}_g|} \sum_{c\in\mathcal{C}_g} \sum_{j=1}^{K} \mathbb{I}\left[c = \arg\max_{c'\in\mathcal{C}_g} T_{m,c'}(x_{cj})\right]. \tag{4}$$

The corresponding semantic weights are

$$\omega_{m,g} = \frac{\exp(r^T_{m,g}/\tau)}{\sum_{\ell\in\mathcal{M}_T} \exp(r^T_{\ell,g}/\tau)}. \tag{5}$$

Thus, $\pi_{m,g}$ and $\omega_{m,g}$ calibrate expert contributions separately for visual-reference and class-name evidence.

3.3 Domain-Conditioned Comparison Geometry

The competence weights determine *how much* each expert contributes. We next use the references to determine *how visual features are compared*, and calibrate the resulting visual branch against class-name evidence.

Domain-conditioned visual geometry. For expert m, we first compute the prototype of each class as $\mu_{m,c} = K^{-1}\sum_{j=1}^{K} u_m(x_{cj})$. We then center every reference feature by its own class prototype and pool the residual variation within domain g:

$$\Sigma_{m,g} = \frac{1}{|\mathcal{C}_g|(K-1)} \sum_{c\in\mathcal{C}_g} \sum_{j=1}^{K} \big(u_m(x_{cj}) - \mu_{m,c}\big)\big(u_m(x_{cj}) - \mu_{m,c}\big)^\top. \tag{6}$$

To stabilize covariance estimation from a small reference library, we use shrinkage toward an isotropic target, a standard covariance regularization strategy (Ledoit & Wolf, 2004). Let $s_g = \sum_m \mathrm{tr}(\Sigma_{m,g})/\sum_m D_m$ denote the domain-level residual variance scale and set $\epsilon = 10^{-6}$. The regularized covariance is

$$H_{m,g} = (1-\lambda)\Sigma_{m,g} + (\lambda s_g + \epsilon) I_{D_m}, \qquad \lambda = 0.5, \tag{7}$$

where I_{D_m} is the $D_m \times D_m$ identity matrix. Its inverse defines the comparison metric, downweighting directions with large within-class variation relative to more stable directions.

The reference-derived geometry calibrates comparisons between image features and visual class prototypes. The semantic branch retains each expert's pretrained image–text similarity, with domain-specific expert contributions determined by $\omega_{m,g}$.

Cross-fitted calibration of the branch weight. The relative usefulness of visual and semantic evidence also varies across domains. We therefore select a domain-specific branch weight w^*_g from $\mathcal{W} = \{0, 0.25, 0.5, 0.75, 1\}$ using nested held-out reference replay.

In outer fold j, the j-th image of every class is held out. We recompute the expert weights $\pi_{m,g}$ and $\omega_{m,g}$, class prototypes $\mu_{m,c}$, and covariance statistics $H_{m,g}$ from the remaining $K-1$ references per class. Visual competence is estimated by an inner held-out replay over these retained references, while semantic competence is evaluated on the retained images. All statistics use the fold-specific reference count; in particular, the prototype and covariance normalizers become $K-1$ and $|\mathcal{C}_g|(K-2)$, respectively.

Using the retained references as the retrieval library, we process each outer-held-out image through Steps 1–3 of Section 3.4 with the fold-specific statistics. For each candidate $w \in \mathcal{W}$, we then evaluate the final composition in Eq. (13) with w in place of $w^*_{\hat{g}(x)}$. We pool the out-of-fold predictions and group them by their inferred coarse-domain context. Within each group g, the candidate with the highest fine-grained top-1 accuracy defines w^*_g. Ties are resolved toward the smaller w. If no held-out image is assigned to a domain, we use the weight selected from all out-of-fold predictions. Finally, we recompute the expert weights, class prototypes, and covariance statistics using the full reference library and cache them together with the selected branch weights.

3.4 QUERY INFERENCE

Given a new query image x, Stage II follows the four steps shown in Figure 2. The visual and semantic branches share the frozen query features and are combined in the final step.

Step 1: retrieval-based coarse-domain context. Each image expert first compares the query with the reference library. For expert m, the retrieval score of class c is

$$a_{m,c}(x) = \mathrm{MeanTop}_3 \left\{ u_m(x)^\top u_m(x_{cj}) : j = 1, \ldots, K \right\}. \tag{8}$$

To place the retrieval scores on comparable scales, we standardize each expert's class-score vector $a_m(x) = (a_{m,c}(x))_{c\in\mathcal{C}}$. For a class-score vector v, define $Z_+(v) = (v - \bar{v})/(\sigma_C(v) + \epsilon)$, where $\bar{v}$ and $\sigma_C(v)$ are its mean and standard deviation over the candidate classes. Each expert produces the distribution $q_m(c \mid x) = \mathrm{softmax}(Z_+(a_m(x)))_c$. The top class of the averaged distributions determines the coarse-domain context:

$$\hat{g}(x) = g\left(\arg\max_{c\in\mathcal{C}} \frac{1}{M} \sum_{m=1}^{M} q_m(c \mid x) \right). \tag{9}$$

The inferred domain indexes the expert weights, comparison geometry, and branch mixture used to score all classes in $\mathcal{C}$.

Step 2: calibrated visual branch. For each expert, we compare the query with class prototypes using the inverse covariance selected by $\hat{g}(x)$. Expanding $-\frac{1}{2}$ times the squared Mahalanobis distance and discarding the class-independent query term yields

$$b_{m,c}(x) = u_m(x)^\top H^{-1}_{m,\hat{g}(x)} \mu_{m,c} - \frac{1}{2} \mu^\top_{m,c} H^{-1}_{m,\hat{g}(x)} \mu_{m,c}. \tag{10}$$

Let $b_m(x) = (b_{m,c}(x))_{c\in\mathcal{C}}$ denote the expert's score vector. Using $Z_+(\cdot)$, the visual branch combines the standardized expert scores into a vector $V(x)$ with entries

$$V_c(x) = \sum_{m=1}^{M} \pi_{m,\hat{g}(x)} Z_+(b_m(x))_c. \tag{11}$$

Step 3: calibrated semantic branch. For text-capable expert m, let $T_m(x) = (T_{m,c}(x))_{c\in\mathcal{C}}$ denote its image–name score vector. We use the stabilized normalization $Z(v) = (v - \bar{v})/\max\{\sigma_C(v), \epsilon\}$ for semantic and branch-level composition. The semantic branch is

$$S(x) = Z\left(\sum_{m\in\mathcal{M}_T} \omega_{m,\hat{g}(x)} Z(T_m(x)) \right). \tag{12}$$

Step 4: final evidence composition. The calibrated branch weight associated with the inferred domain combines the visual and semantic evidence:

$$F_c(x) = \left(1 - w^*_{\hat{g}(x)}\right) Z(V(x))_c + w^*_{\hat{g}(x)} S_c(x). \tag{13}$$

The final prediction is $\hat{y}(x) = \arg\max_{c\in\mathcal{C}} F_c(x)$.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmarks. We evaluate three joint recognition tasks with five labeled references per class. CUB-200-2011, Stanford Cars, FGVC-Aircraft, and Oxford Flowers-102 form a 598-class benchmark with 23,317 queries (Wah et al., 2011; Krause et al., 2013; Maji et al., 2013; Nilsback & Zisserman, 2008). The query domain is unknown at inference. We separately evaluate all 10,000 iNaturalist 2021 species using 100,000 validation queries (Visipedia, 2021).

To further examine recognition of specific product models and designs among man-made objects, we evaluate product-identity recognition on Stanford Online Products (SOP) (Song et al., 2016).

Table 1: **Fine-grained recognition across man-made and natural domains.** Top-1 accuracy (%). Venue/year entries cite the source algorithms; variant labels refer to the evaluated implementations.

		Man-made domains			Natural domains		
Method	Venue / Year	Aircraft	Cars	SOP	CUB	Flowers	iNat-10k
Individual frozen backbones							
DINOv3	arXiv 2025	49.81	83.61	60.81	84.74	99.71	48.21
BioCLIP2	NeurIPS 2025	35.22	83.12	65.80	92.65	98.68	85.97
SigLIP2	arXiv 2025	61.78	89.88	75.85	66.43	98.16	24.65
FG-CLIP2	ICML 2026	60.04	86.42	75.98	67.26	98.46	29.67
Generative assistance							
UniFGVC	arXiv 2025	61.06	94.63	–	–	96.34	–
Additional selection controls							
Global LogME	ICML 2021	45.10	79.52	–	74.37	99.71	–
Global PARC	NeurIPS 2021	49.81	83.61	–	84.74	99.71	–
Reference-routed classifiers							
LogME	ICML 2021	45.10	79.52	46.22	87.02	99.71	41.25
PARC	NeurIPS 2021	49.81	89.88	75.93	92.56	99.71	85.97
Tip-Adapter	ECCV 2022	52.90	93.37	67.77	91.39	85.22	88.65
APE	ICCV 2023	58.93	93.19	–	92.42	89.90	89.34
TIP-X	ICCV 2023	53.80	93.42	–	91.47	85.66	88.73
GDA	ICLR 2024	68.53	92.94	84.03	93.10	99.20	88.74
ECALP	ICLR 2025	55.63	88.31	–	91.25	97.76	–
TextClassifier	This work	48.70	93.00	–	91.22	83.53	88.43
Multi-model methods							
KNORA-U	Pattern Recogn. 2008	57.37	87.19	70.56	84.48	99.07	79.02
KNORA-E	Pattern Recogn. 2008	55.96	85.35	69.49	85.00	99.38	75.52
ZS-En	ICML 2024	43.20	92.25	77.47	91.58	82.76	88.49
TF-En	ICML 2024	46.42	93.10	77.84	91.18	83.28	88.51
COSMIC-fixed	CVPR 2025	14.04	73.26	–	91.61	79.90	–
SOTA Transductive	CVPR 2026	52.12	94.53	–	92.04	91.32	–
SOTA Inductive	CVPR 2026	49.99	94.44	–	91.75	85.72	–
Our method							
Ours (visual only)	This work	73.51	93.11	87.57	92.86	99.72	85.42
Ours	This work	**73.51**	**95.05**	**87.57**	**93.23**	**99.72**	**89.46**

This requires finer distinctions than broad product categories: for example, two chairs in the same coarse category, or two visually similar bicycles, remain distinct recognition targets. We retain 4,431 identities with at least six images from the official test split; five images per identity serve as references and the remaining 15,732 images as queries. Each task uses its complete candidate vocabulary.

Experts and calibration. We use frozen DINOv3 ViT-L/16, BioCLIP2 ViT-L/14, FG-CLIP2-base, and SigLIP2-base-patch16-224 encoders. The latter three also provide text encoders. Expert reliability, comparison geometry, and branch weights are calibrated from references as described in Section 3; query labels are used only for evaluation. SOP uses visual evidence only because usable identity-level class names are unavailable. We therefore omit the semantic branch and set $w_g^* = 0$ for every SOP domain, so that the final score reduces to $F_c(x) = Z(V(x))_c$.

Comparisons and metrics. We report Top-1 accuracy (%) for comparison methods. Except for the generative-assistance methods, baselines are re-evaluated using the same available backbone pool, candidate vocabulary, reference construction, and queries wherever applicable. The source methods are reviewed in Section 2 and linked through the venue/year entries in Table 1. For methods that do not natively compose multiple backbones, we construct domain-conditioned single-expert baselines. A backbone is selected for each domain using a method-specific reference-only criterion. At inference, the predicted coarse-domain context determines which selected backbone is used to score the complete candidate vocabulary. For UniFGVC, which uses an additional VLM, we report the results published in their original papers. SOTA Transductive uses unlabeled queries, SOTA Inductive uses a separate unlabeled adaptation set. These resource differences are retained in the comparisons. Further implementation details are provided in the appendix.

4.2 Main Results

Four-domain recognition. Table 1 shows complementary domain profiles: BioCLIP2 is the strongest individual backbone on CUB, whereas SigLIP2 leads on Cars and Aircraft. Our method

Table 2: **Product identities without semantic text.** Top-1 accuracy (%) using the same reference seed.

Method	Bicycle	Cabinet	Chair	Coffee maker	Fan	Kettle	Lamp	Mug	Sofa	Stapler	Table	Toaster	Overall	Domain avg.
I. Individual backbones														
DINOv3-L/16	26.72	56.67	70.40	39.68	64.56	58.31	74.54	73.62	64.86	65.24	80.15	59.87	60.81	61.22
BioCLIP2	52.53	63.43	73.23	42.76	67.16	55.02	75.78	76.11	76.25	66.34	79.65	58.68	65.80	65.58
FG-CLIP2	69.68	74.55	82.51	53.38	76.95	68.59	83.93	83.01	79.72	77.85	86.41	70.23	75.98	75.57
SigLIP2	67.91	74.46	81.15	54.61	78.19	67.95	84.88	81.70	81.24	74.42	87.64	72.49	75.85	75.55
II. Reference-based backbone selection														
Tip-Adapter	61.99	63.77	74.48	42.14	66.17	55.26	78.26	79.97	72.56	68.05	81.80	62.46	67.77	67.24
GDA	79.50	85.24	87.15	69.71	83.15	81.85	89.31	86.40	87.20	85.68	89.37	79.61	84.03	83.68
LogME	16.05	40.30	51.56	23.27	46.59	34.78	63.62	69.54	51.19	45.66	68.04	44.88	46.22	46.29
PARC	67.97	74.46	81.15	54.61	78.07	67.95	83.99	81.70	81.24	77.85	87.56	72.60	75.93	75.76
III. Multi-model fusion														
ZS-En	71.39	75.64	82.51	56.28	77.08	69.88	85.17	84.12	82.76	80.42	87.81	73.57	77.47	77.22
TF-En	71.45	76.35	82.91	56.72	78.07	70.76	85.47	84.05	82.86	80.29	88.30	73.68	77.84	77.58
KNORA-U	60.34	68.17	77.53	48.82	70.01	60.40	80.15	77.49	80.26	72.22	83.36	65.48	70.56	70.35
KNORA-E	58.51	67.83	77.08	47.06	69.15	59.04	78.85	78.38	77.55	69.77	82.37	63.97	69.49	69.13
SOTA (support-init)	66.08	75.64	82.68	55.58	78.69	70.76	84.11	83.43	81.89	81.76	88.30	73.79	76.93	76.89
Ours (visual only)	**85.97**	**87.04**	**88.68**	**78.93**	**85.75**	**85.14**	**89.72**	**88.95**	**89.37**	**92.66**	**91.93**	**86.84**	**87.57**	**87.58**

Table 3: **Component ablations on the four-domain task.** Top-1 accuracy (%), aggregated over matched reference splits with all 598 candidates. Parentheses show percentage-point differences from the full method: red + indicates higher accuracy and green − lower or equal accuracy.

Configuration	Aircraft	Cars	CUB	Flowers
Visual only	72.73 (-0.78)	92.99 (-2.06)	92.80 (-0.43)	99.72 (-0.00)
Semantic only	48.25 (-25.26)	93.94 (-1.11)	91.35 (-1.88)	83.51 (-16.21)
Fixed branch weight 0.5	61.90 (-11.61)	94.98 (-0.07)	92.76 (-0.47)	97.82 (-1.90)
Global branch weight	68.14 (-5.37)	94.96 (-0.09)	93.21 (-0.02)	99.73 (+0.01)
Isotropic geometry	62.99 (-10.52)	94.79 (-0.26)	93.01 (-0.22)	99.72 (-0.00)
Global covariance	71.18 (-2.33)	94.97 (-0.08)	93.09 (-0.14)	99.73 (+0.01)
Equal visual weights	71.24 (-2.27)	94.84 (-0.21)	92.92 (-0.31)	99.72 (-0.00)
Full method	**73.51**	**95.05**	**93.23**	**99.72**

attains the highest displayed accuracy across the four domains. Its advantage over GDA is largest on Aircraft (73.507% versus 68.527%, a 4.980-point gain), while the CUB gap is 0.138 points and Flowers is already near saturation.

Large-scale species recognition. On iNat-10k, our method reaches 89.457%, closely followed by APE at 89.343%. The visual-only variant obtains 85.415%, below the BioCLIP2 prototype at 85.967%; the complete result therefore benefits from combining visual references with class-name evidence.

Product identity recognition. Table 2 further reports SOP results across 12 coarse product groups, while every query is still classified against all 4,431 product identities. Without identity-level text, our visual branch reaches 87.567% overall, exceeding the GDA visual adaptation by 3.541 points and obtaining the highest displayed accuracy across the reported product groups. This shows that reference-calibrated expert composition remains effective for fine-grained product identities even without class-name semantics.

4.3 Ablation and Robustness Analysis

Table 3 evaluates the main components using the same expert pool, reference splits, queries, and 598 candidates. When geometry or visual weights are changed, the branch mixture is recalibrated using the same reference-only procedure.

Evidence calibration. Removing either visual or semantic evidence generally degrades performance, while fixed or globally shared branch weights are notably less effective on domains such as Aircraft. This confirms that the contribution of semantic evidence should be calibrated according to domain-specific reference statistics rather than applied uniformly.

(a) Domain-dependent feature discriminability (b) Model-dependent regional evidence

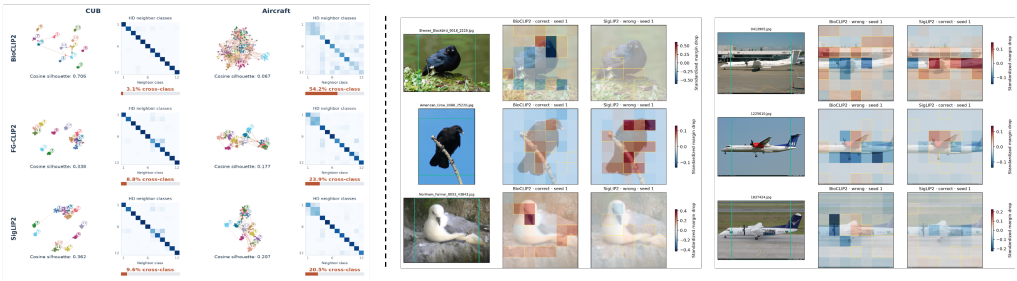


Figure 3: **Understanding domain expertise through feature discriminability and regional evidence.** (a) Domain-dependent feature discriminability. (b) Model-dependent regional evidence.

Geometry and expert weighting. Replacing the domain-conditioned geometry with isotropic or global variants, or assigning equal weights to visual experts, generally reduces recognition accuracy, with the clearest effects appearing on Aircraft. This supports both reference-derived feature geometry and domain-specific expert weighting. On CUB, Cars, and especially Flowers, several variants differ only marginally from the full method, and a few are slightly higher. These datasets already operate at high accuracy, leaving limited headroom, so small fluctuations are expected and do not change the overall ablation trend. Further ablation analyses and experimental settings are provided in the appendix.

4.4 Understanding Domain Expertise

Figure 3 provides two complementary views of domain expertise: how well each representation separates fine-grained classes, and which image regions different experts rely on for discrimination.

(a) Domain-dependent feature discriminability. Figure 3(a) compares BioCLIP2, FG-CLIP2, and SigLIP2 on matched CUB and Aircraft subsets using t-SNE projections (van der Maaten & Hinton, 2008) alongside original-space neighborhood statistics. The relative feature separation of the three experts changes substantially across domains: BioCLIP2 produces more discriminative features on CUB, whereas FG-CLIP2 and SigLIP2 separate Aircraft classes more clearly. This domain-dependent ordering motivates estimating expert reliability from reference features rather than assuming a universally strongest representation.

(b) Model-dependent regional evidence. The occlusion-sensitivity maps (Zeiler & Fergus, 2014) in Figure 3(b) show that BioCLIP2 and SigLIP2 often rely on different regions of the same image. Their different pretraining objectives and data distributions (Gu et al., 2025; Tschannen et al., 2025) are plausible sources of these distinct visual dependencies, although the visualizations do not isolate their causal effects. Thus, expert complementarity appears not only in feature-space separability but also in the visual evidence used for discrimination.

Together, these observations motivate our reference-calibrated expert composition: domain expertise is estimated from reference features, and complementary frozen experts are softly combined instead of relying on a single pretrained representation.

5 Conclusion

We presented Reference-Calibrated Expert Composition for open-domain fine-grained recognition. A labeled reference library calibrates expert contributions, domain-conditioned comparison geometry, and the balance between visual and semantic evidence, while all pretrained encoders remain frozen. Experiments on mixed-domain categories, species, and product identities support the effectiveness of this approach. The framework requires reference coverage and a predefined class-to-domain mapping, and its computational cost depends on the expert pool and candidate vocabulary.

AI USE STATEMENT

Generative AI tools were used solely for language editing and polishing. They were not used to generate research ideas, methods, experimental results, or scientific claims. All AI-assisted edits were reviewed by the authors for accuracy and consistency. The authors take full responsibility for the final content of this work.

REFERENCES

Daniel Bolya, Rohit Mittapalli, and Judy Hoffman. Scalable diverse model selection for accessible transfer learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 19301–19312, 2021. URL https://arxiv.org/abs/2111.06977.

Jianyang Gu, Samuel Stevens, Elizabeth G. Campolongo, Matthew J. Thompson, Net Zhang, Jiaman Wu, Andrei Kopanev, Zheda Mai, Alexander E. White, James Balhoff, Wasila Dahdul, Daniel Rubenstein, Hilmar Lapp, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. BioCLIP 2: Emergent properties from scaling hierarchical contrastive learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2505.23883.

Hongyu Guo, Xiangzhao Hao, Jiarui Guo, Haiyun Guo, Jinqiao Wang, and Tat-Seng Chua. UniFGVC: Universal training-free few-shot fine-grained vision classification via attribute-aware multimodal retrieval. arXiv preprint arXiv:2508.04136, 2025. URL https://arxiv.org/abs/2508.04136. Version 3, revised February 2026.

Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. TransFG: A transformer architecture for fine-grained recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 852–860, 2022. doi: 10.1609/aaai.v36i1.19967. URL https://ojs.aaai.org/index.php/AAAI/article/view/19967.

Zhanxuan Hu, Qiyu Xu, Yu Duan, Yonghang Tai, and Huafeng Li. SOTA: Self-adaptive optimal transport for zero-shot classification with multiple foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://arxiv.org/abs/2506.13723.

Fanding Huang, Jingyan Jiang, Qinting Jiang, Hebei Li, Faisal Nadeem Khan, and Zhi Wang. COSMIC: Clique-oriented semantic multi-space integration for robust CLIP test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2503.23388.

Albert H. R. Ko, Robert Sabourin, and Alceu Souza Britto, Jr. From dynamic classifier selection to dynamic ensemble selection. *Pattern Recognition*, 41(5):1718–1731, 2008. URL https://www.sciencedirect.com/science/article/pii/S0031320307004499.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 554–561, 2013. doi: 10.1109/ICCVW.2013.77. URL https://vision.stanford.edu/pdf/3drr13.pdf.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2202.10054.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4. URL https://www.ledoit.net/Well-conditioned2004.pdf.

Yushu Li, Yongyi Su, Adam Goodge, Kui Jia, and Xun Xu. Efficient and context-aware label propagation for zero-/few-shot training-free adaptation of vision-language model. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2412.18303.

Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1449–1457, 2015. URL https://openaccess.thecvf.com/content_iccv_2015/html/Lin_Bilinear_CNN_Models_ICCV_2015_paper.html.

Zhihe Lu, Jiawang Bai, Xin Li, Zeyu Xiao, and Xinchao Wang. Beyond sole strength: Customized ensembles for generalized vision-language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2311.17091.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151, 2013. URL https://arxiv.org/abs/1306.5151.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, pp. 722–729, 2008. doi: 10.1109/ICVGIP.2008.47. URL https://www.robots.ox.ac.uk/~vgg/publications/2008/Nilsback08/.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023. URL https://arxiv.org/abs/2304.07193.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi: 10.1016/0377-0427(87)90125-7. URL https://www.sciencedirect.com/science/article/pii/0377042787901257.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. arXiv preprint arXiv:2508.10104, 2025. URL https://arxiv.org/abs/2508.10104.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4004–4012, 2016. URL https://cvgl.stanford.edu/projects/lifted_struct/.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786, 2025. URL https://arxiv.org/abs/2502.14786.

Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. SuS-X: Training-free name-only transfer of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2725–2736, 2023. URL https://arxiv.org/abs/2211.16198.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL https://jmlr.org/papers/v9/vandermaaten08a.html.

Visipedia. iNaturalist 2021 competition dataset. Official dataset documentation, 2021. URL https://github.com/visipedia/inat_comp/tree/master/2021.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. URL https://www.vision.caltech.edu/datasets/cub_200_2011/.

Zhengbo Wang, Jian Liang, Lijun Sheng, Ran He, Zilei Wang, and Tieniu Tan. A hard-to-beat baseline for training-free CLIP-based adaptation. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2402.04087.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://arxiv.org/abs/2109.01903.

Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Ji Ao, Dawei Leng, and Yuhui Yin. FG-CLIP 2: A bilingual fine-grained vision-language alignment model. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2510.10921.

Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Ze_Yang_Learning_to_Navigate_ECCV_2018_paper.html.

Zhipeng Ye, Feng Jiang, Qiufeng Wang, Kaizhu Huang, and Jiaqi Huang. IDEA: Image description enhanced CLIP-adapter for image classification. *Pattern Recognition*, 171:112224, 2026. doi: 10.1016/j.patcog.2025.112224. URL https://www.sciencedirect.com/science/article/pii/S0031320325008854.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12133–12143, 2021. URL https://proceedings.mlr.press/v139/you21b.html.

Hong-Tao Yu, Chen-Wei Xie, Yuxin Peng, Serge Belongie, and Xiu-Shen Wei. Benchmarking large vision-language models on fine-grained image tasks: From evaluation to diagnosis. arXiv preprint arXiv:2606.19053, 2026. URL https://arxiv.org/abs/2606.19053.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 818–833, 2014. doi: 10.1007/978-3-319-10590-1_53. URL https://arxiv.org/abs/1311.2901.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-Adapter: Training-free adaption of CLIP for few-shot classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 493–510, 2022. URL https://arxiv.org/abs/2207.09519.

Didi Zhu, Zhongyi Sun, Zexi Li, Tao Shen, Ke Yan, Shouhong Ding, Kun Kuang, and Chao Wu. Model tailor: Mitigating catastrophic forgetting in multi-modal large language models. arXiv preprint arXiv:2402.12048, 2024. URL https://arxiv.org/abs/2402.12048.

Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. Not all features matter: Enhancing few-shot CLIP with adaptive prior refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2605–2615, 2023. URL https://arxiv.org/abs/2304.01195.

Table 4: **Joint recognition of 10,000 species.** Top-1 accuracy (%).

Method	Plants	Mammals	Amph.	Rept.	Other animals	Arachn.	Birds	Fungi	Insects	Mollusks	Ray-finned fishes	Overall	Domain avg.
Individual frozen backbones													
DINOv3	58.74	57.03	31.12	41.60	57.18	43.46	50.31	57.80	27.99	47.16	60.05	48.21	48.40
BioCLIP2	85.78	82.07	71.35	75.43	84.86	86.14	83.41	87.30	90.56	86.33	82.40	85.97	83.24
SigLIP2	27.50	36.95	11.71	18.02	31.48	21.50	22.42	38.24	18.96	29.29	29.07	24.65	25.92
FG-CLIP2	33.71	41.71	15.00	20.16	37.04	25.36	27.96	43.70	22.23	34.14	33.11	29.67	30.37
Reference-routed classifiers													
LogME	27.04	52.32	14.71	46.45	38.38	25.10	48.63	68.68	59.25	29.53	40.60	41.25	40.97
PARC	85.78	82.07	71.35	75.43	84.86	86.14	83.41	87.30	90.56	86.33	82.40	85.97	83.24
Tip-Adapter	89.30	83.74	77.65	83.77	89.93	90.26	86.39	88.01	90.68	87.75	88.63	88.65	86.92
TIP-X	89.31	84.02	77.71	83.74	89.79	90.39	86.49	88.18	90.85	87.81	88.80	88.73	87.01
GDA	89.37	83.82	77.82	83.67	89.93	90.20	86.44	88.30	90.80	88.05	88.74	88.74	87.01
ECALP	–	–	–	–	–	–	–	–	–	–	–	–	–
TextClassifier	89.11	83.33	77.24	83.64	89.58	90.07	86.14	87.83	90.43	87.87	88.42	88.43	86.70
Multi-model methods													
KNORA-U	77.95	71.42	63.29	69.30	75.21	80.26	75.81	79.91	86.37	76.45	72.46	79.02	75.31
KNORA-E	75.00	72.07	51.35	61.69	73.87	76.60	71.58	79.50	82.25	74.20	71.58	75.52	71.79
ZS-En	89.16	83.46	77.59	**84.09**	89.51	**90.59**	86.22	87.54	90.48	87.57	88.47	88.49	86.79
TF-En	89.13	83.66	77.65	84.06	89.58	90.46	86.30	87.65	90.52	87.69	88.31	88.51	86.82
Our method													
Ours (visual only)	85.33	80.98	70.35	76.23	83.94	85.10	82.91	87.04	89.75	84.56	82.79	85.42	82.63
Ours	**89.77**	**84.19**	**78.29**	83.99	**90.00**	90.46	**87.03**	**89.44**	**92.25**	**89.29**	**89.18**	**89.46**	**87.63**

A Complete iNaturalist Results

Evaluation protocol. Table 4 expands the iNat-10k results in Table 1. All 10,000 species remain candidates for each of the 100,000 queries. The 11 taxonomic groups define the coarse contexts for calibration and reporting. Reference-dependent results use seed 1, while ZS-En retains its fixed single-run result. Overall accuracy weights queries equally, whereas Domain avg. weights the 11 groups equally.

Results across taxonomic groups. Our method achieves the highest overall accuracy (89.46%) and domain average (87.63%) among the compared methods. It also obtains the highest accuracy in nine of the eleven taxonomic groups, with ZS-En leading only on reptiles and arachnids. Compared with the visual-only variant, the complete method improves consistently across all groups, showing that class-name evidence complements reference-based visual composition even in a 10,000-species vocabulary.

B Additional Implementation Details

Global backbone selection. Global LogME and Global PARC in Table 1 select one backbone for the entire joint task. LogME ranks the four backbones by Bayesian evidence on the pooled labeled references and predicts with the selected Bayesian linear classifier. PARC ranks them by the agreement between feature similarities and label similarities, using a fixed reference probe of up to 500 images and its 32-dimensional projection; prediction then uses cosine similarity to class-mean reference prototypes. The selected backbone is shared by all queries, without domain-conditioned routing.

Reference-routed classifiers. The routed LogME and PARC variants apply their selection criteria separately within each reference domain. Other routed classifiers select backbones by five-fold reference-only validation. Methods requiring class-name text select among BioCLIP2, FG-CLIP2, and SigLIP2; visual criteria can also select DINOv3. Query routing averages the four experts' standardized Top-3 reference probabilities. The predicted domain selects the classifier, while all candidate classes remain available.

Multi-model and product-identity settings. KNORA-U and KNORA-E use three references per class for their prototype pool and two for dynamic-selection calibration. SOTA Transductive uses unlabeled queries, whereas SOTA Inductive uses separate unlabeled adaptation data. On SOP, GDA and Tip-Adapter use their visual adaptations, ZS-En and TF-En use visual prototypes, and SOTA (support-init) initializes classes from reference prototypes. These variants do not require identity-level class names.

Table 5: **Visual expert composition.** Top-1 accuracy (%), aggregated over matched reference splits with all 598 candidates. The semantic branch is disabled.

Visual experts	Aircraft	Cars	CUB	Flowers
Single expert				
DINOv3	66.62	87.12	85.66	99.66
BioCLIP2	37.52	84.54	92.87	99.27
FG-CLIP2	68.93	92.65	75.47	99.18
SigLIP2	65.43	92.27	70.94	98.86
Two experts				
DINOv3, BioCLIP2	67.48	89.40	92.77	99.71
DINOv3, FG-CLIP2	73.52	91.74	85.65	99.67
DINOv3, SigLIP2	71.52	92.64	85.58	99.67
BioCLIP2, FG-CLIP2	68.19	91.71	92.88	99.53
BioCLIP2, SigLIP2	65.33	92.28	92.86	99.55
FG-CLIP2, SigLIP2	69.48	93.10	74.93	99.21
Three experts				
DINOv3, BioCLIP2, FG-CLIP2	72.99	91.85	92.80	99.72
DINOv3, BioCLIP2, SigLIP2	71.42	92.55	92.80	99.72
DINOv3, FG-CLIP2, SigLIP2	72.91	93.10	85.71	99.69
BioCLIP2, FG-CLIP2, SigLIP2	69.27	92.90	92.91	99.57
Full visual pool				
DINOv3, BioCLIP2, FG-CLIP2, SigLIP2	72.73	92.99	92.80	99.72

C Additional Ablation and Robustness Analysis

Component settings. We retain the four-domain benchmark, 23,317 queries, and all 598 candidates used in Section 4.3. Visual only and Semantic only retain one evidence branch. Fixed branch weight 0.5 assigns equal weight to the two standardized branch scores. Global branch weight selects one mixture for all domains using the same reference-only validation procedure. Isotropic geometry replaces each domain covariance by a scaled identity with the shared scale; Global covariance pools residuals across all reference classes for each expert. Equal visual weights averages the four visual experts. Geometry and expert-weight interventions recalibrate the branch mixture from references.

Expert composition. Table 5 evaluates all 15 nonempty visual expert subsets. The semantic branch is disabled, and each subset recalibrates its router, reliability weights, and shared covariance scale. Single-expert rows therefore retain our covariance-based classifier, unlike the cosine-prototype baselines in Table 1. Results are aggregated over the same reference splits as the main component ablations; they are not the seed-1 main-table results.

Effect of removing an expert. Removing DINOv3 lowers Aircraft accuracy from 72.73% to 69.27%, while removing BioCLIP2 lowers CUB accuracy from 92.80% to 85.71%. Some smaller subsets slightly exceed the full pool on individual domains. This supports complementary domain expertise, rather than a monotonic benefit from adding every expert. Because routing and calibration are recomputed for each subset, the table evaluates the resulting systems rather than an isolated change to one score.

Number of references. Table 6 evaluates 3, 5, 9, and 15 references per class while retaining the same 23,317 queries and all 598 candidate classes. CUB, Cars, and Aircraft use only official training images. Flowers has only ten official training images per class, so its 15-reference setting uses all ten training images and five images sampled from the official validation split. No test query is used as a reference.

Robustness across domains. Aircraft shows the largest benefit from additional references, improving from 67.62% with three references to 73.51% with five, 77.54% with nine, and 80.69% with fifteen. The 15-reference result exceeds the nine-reference setting by 3.15 points and the five-reference baseline by 7.18 points. Cars and CUB improve more gradually as the reference budget increases, whereas Flowers remains essentially saturated at 99.72%. These results indicate that richer reference evidence is most beneficial when the underlying task still has substantial recognition headroom. Because the 15-reference setting expands the Flowers reference pool and recalibrates the joint sys-

Table 6: **Effect of reference count.** The full method is recalibrated for each reference count. All rows use the same test queries and 598 candidates. The 15-reference row expands the Flowers reference pool to ten training and five validation images per class; the other domains retain training-only references.

References per class	Aircraft	Cars	CUB	Flowers
3	67.62	94.51	92.62	99.71
5	73.51	95.05	93.23	99.72
9	77.54	95.44	93.59	99.72
15	80.69	95.55	93.79	99.72

Table 7: **Comparison with published training-free results.** Top-1 accuracy (%). Literature rows retain their published settings.

Method	Setting	Backbone	Aircraft	Cars	Flowers	Source
Original-paper results						
TIP-X	16-shot	CLIP RN50	29.61	–	90.54	Table 16
GDA	16-shot	CLIP RN50	40.61	75.12	95.72	Table 1
GDA	16-shot	CLIP ViT-B/16	50.33	82.63	97.92	Table 9
ECALP	Zero-shot	CLIP ViT-B/16	29.49	68.20	75.96	Table 1
COSMIC	Zero-shot TTA	CLIP ViT-B/16+DINOv2	31.44	71.31	82.14	Table 2
ZS-En	Zero-shot	Four CLIP encoders	25.68	70.76	73.16	Table 1
SOTA Transductive	Zero-shot, transductive	CLIP ViT-B/16+DINOv3	33.00	79.60	84.20	Table 1
SOTA Inductive	Zero-shot, inductive	CLIP ViT-B/16+DINOv3	31.80	78.80	84.10	Table 4
Baseline results reported in GDA						
Tip-Adapter	16-shot	CLIP RN50	29.76	66.77	89.89	GDA Table 1
APE	16-shot	CLIP RN50	31.23	70.45	91.96	GDA Table 1
Our method						
Ours	5-shot, joint task	Four frozen experts	**73.51**	**95.05**	**99.72**	This work

tem, it should be interpreted as an extended-reference experiment rather than a strictly controlled same-pool comparison. The reported values are mean accuracies rather than claims of statistical significance.

D Comparison with Published Results

Comparison protocol. Table 7 compares our result with published training-free configurations on Aircraft, Cars, and Flowers. The literature evaluates individual datasets under its own settings; our method uses five references per class in the joint 598-class task. Results verified in the corresponding papers and baseline values transcribed by GDA are grouped separately.

Published-result comparison. Our method achieves the highest displayed accuracy on all three datasets among the listed configurations. Relative to GDA with CLIP ViT-B/16, the strongest listed literature result in each column, the gaps are 23.18 points on Aircraft, 12.42 on Cars, and 1.80 on Flowers. These are differences between published systems: backbone resources, reference budgets, unlabeled-data access, candidate vocabularies, and test splits differ. They therefore complement the local evaluations in Table 1, rather than isolate an algorithmic gain.

Result sources. TIP-X uses its own Table 16, which provides Aircraft and Flowers but no Cars result. Tip-Adapter and APE are taken from GDA Table 1 and are identified as secondary-source values. The remaining rows use the indicated tables in their corresponding papers; the SOTA rows keep the same CLIP ViT-B/16 and DINOv3 combination. Only classification accuracy is included, excluding supervised linear probes, base-to-new harmonic means, and retrieval Recall@K.

Table 8: **Inference efficiency.** Time/image denotes amortized batch inference time. For cached-feature methods, memory is the PyTorch peak allocated memory. The Qwen memory value (†) reports observed process GPU occupancy and is therefore not directly comparable to PyTorch allocation.

Method	Total (s)	ms/image	Allocated (MiB)
Cached-feature routing and classification			
Tip-Adapter	0.234	0.010	397.6
APE	0.254	0.011	429.4
GDA	0.198	0.008	377.0
ECALP	131.117	5.623	457.8
Ours	1.163	0.050	428.2
Image-input Top-5 reranking			
Qwen3-VL-4B-Instruct	148.214	44.469	38108†

E Inference Efficiency and Resource Usage

Measurement protocol. Table 8 reports inference measurements on a single NVIDIA L40. The cached-feature methods are evaluated on all 23,317 queries from the four-domain benchmark using seed 1. Each method runs in a separate process, with one complete warmup followed by three timed passes; synchronized wall-clock times are summarized by their median. Qwen3-VL-4B-Instruct is measured for the corresponding four-domain Top-5 reranking stage over two completed passes. GPU monitoring confirmed that no competing compute process was active during measurement.

Cached-feature classification. Our method processes the complete four-domain query set in 1.163 seconds, corresponding to 0.050 ms per image, with 428.2 MiB peak allocated memory. The multi-expert routing and composition introduce some additional overhead relative to lightweight single-model baselines such as Tip-Adapter and GDA, but the absolute cost remains small. In particular, the classifier requires well below 1 GiB of additional GPU memory and remains practical on resource-constrained hardware once expert features are available. Its classification cost is also substantially lower than the more expensive sequential ECALP procedure.

Qwen reranking. Qwen3-VL-4B-Instruct uses BF16, batch size 16, and one output token constrained to the Top-5 retrieved candidates. Image loading, preprocessing, visual encoding, and answer generation are timed, while model loading and candidate retrieval are excluded. One batch is used for warmup, and image-processing and prefix caches are disabled. The measured process occupies approximately 38 GiB of GPU memory, reflecting the substantially larger resource footprint of introducing a generative vision-language model at inference time.

Interpretation. Overall, reference-calibrated expert composition does not introduce a large inference burden beyond conventional cached-feature classifiers. Although it is moderately more expensive than the lightest single-model methods, its runtime and memory requirements remain modest and are substantially below those of either more computationally involved adaptation or large-model reranking. Importantly, the first block measures classification over cached expert features and therefore excludes image encoding, encoder weights, feature loading, and reference calibration. The Qwen block additionally includes image processing and visual-language inference, and its memory value includes model weights, CUDA context, and preallocated KV cache. The two blocks should therefore be interpreted as stage-level resource measurements rather than a matched end-to-end speed or memory comparison.