NEURAL CATALOG: SCALING SPECIES RECOGNITION WITH CATALOG OF LIFE-AUGMENTED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling fine-grained recognition from curated benchmarks (200 species) to real-world inventories (11,202 species) costs vision-language models more than 30 points of accuracy, exposing a structural limitation: discriminating among thousands of visually similar candidates requires reasoning over *relative* differences rather than absolute matching. We introduce CoLAG (Catalogue of Life Augmented Generation), a training-free system built on a catalogue of 11,202 species indexed as *discriminative summaries* rather than raw articles. CoLAG pairs a multimodal retrieval stage with neighborhood-subtracted embeddings that amplify discriminative dimensions by suppressing features shared among confusable species, and a comparative reasoning stage that presents a multimodal LLM with retrieved candidates, discriminative descriptions, and calibrated retrieval scores for joint evaluation. Across five bird benchmarks, marine species, and Pokémon, CoLAG adds 35.1 points on average over direct prompting across six MLLM backbones, and 10.0 points over independent scoring.

1 INTRODUCTION

The American Museum of Natural History estimates 18,043 bird species Barrowclough et al. (2016), while CUB-200-2011 Wah et al. (2011), the standard benchmark for bird classification, contains 200 classes. This gap exposes a limitation of current evaluation protocols: they do not reflect the open-vocabulary nature of real-world recognition, where the candidate set is large, evolving, and often unbounded.

Scaling vision-language models (VLMs) to the full species inventory exposes a failure that closed-set benchmarks systematically hide. As shown in Table 1, expanding the candidate set from a curated benchmark to 11,202 bird species costs over 30 points of accuracy across CLIP, OpenCLIP, and SigLIP. The collapse is driven by the density of visually similar distractors in the full label space. Distinguishing a Cooper's Hawk from a Sharp-shinned Hawk, or a Chilean Hawk from a Bicolored Hawk, requires reasoning over fine-grained morphology, beak curvature, plumage barring, tail proportions, which VLMs aligned on holistic image-caption pairs cannot perform at scale.

Table 1: Average accuracy drop when moving from a closed-world setting (species present in dataset) to an open-world setting with all 11,202 species.

Model	CUB		Birdsnap		
	Bounded	Open	Bounded	Open	Average Drop
CLIP	62.9	24.6	51.3	26.5	31.5
OpenCLIP	74.3	35.2	61.4	33.7	33.4
SigLIP	77.2	49.9	64.6	43.6	24.1

A natural approach for unseen classes is to leverage textual descriptions, as pioneered by Elhoseiny et al. (2013). Descriptions are readily available, scalable, and capture visually relevant attributes without labeled training data. However, sources such as Wikipedia are not written for visual discrimination: taxonomy, etymology, historical context, and conservation status dilute the attributes a model must compare, so feeding raw text into a recognition pipeline increases context length without a proportional increase in visual signal.

A simple way to exploit such a corpus is a standard retrieval-augmented pipeline: embed the articles, take the cosine nearest neighbours of the query, and hand them to a multimodal LLM. In our open-vocabulary setting this is not enough. Instantiated with a strong retriever it lifts average accuracy from 30.5% to only 35.0% on Qwen3-VL, and baseline multimodal RAG systems reach at most

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.

30.6% (Tables 3 and 4). What closes the remaining gap is not retrieval augmentation as such, but three specific departures from that naive recipe, one at each stage of the pipeline: *what* we index, *how* we score candidates, and *how* the final decision is made.

We propose CoLAG (Catalogue of Life Augmented Generation), which reformulates open-vocabulary recognition as comparative reasoning over a small set of high-quality candidate hypotheses. CoLAG indexes Wikipedia articles for 11,202 bird species into a structured catalogue accessible to any recognition model. What we index are distilled summaries rather than raw articles: as part of indexing, each article is rewritten into a concise entry retaining only visually discriminative attributes. This cuts the average entry from 552 to 127 words, over 75%, and *raises* recognition accuracy by 2.2 points over feeding the raw articles. CoLAG then leverages this catalogue through its two recognition modules.

$CoLAG_{ret}$ is a multimodal retrieval pipeline that fuses textual and visual embeddings during ranking and refines candidates with a dedicated visual re-ranking stage, so the final candidate set is both semantically and visually consistent with the query. We score candidates using a corrected similarity rather than a raw cosine. At the scale of $11{,}202$ confusable classes the fused embedding space is dominated by a shared coarse-category direction, and a few species act as *hubs* that sit close to everything and so surface for queries they are unrelated to. $CoLAG_{ret}$ corrects this in the space where it lives, down-weighting candidates whose neighborhood is crowded (a hubness penalty) and high-pass filtering every database *and* query embedding by subtracting its own local neighborhood centroid (symmetric whitening).

$CoLAG_{gen}$ then reasons over the retrieved candidates with a multimodal large language model (MLLM). The decision is made by one joint comparison rather than k independent judgments. Rather than predicting from an unconstrained label space, the MLLM sees the whole candidate set, each entry paired with a discriminative summary and a retrieval score, and is asked to name the best match. The naive alternative, scores each retrieved candidate in its own context and takes the $\arg\max$; that costs 10.0 points on average, because scores formed in separate contexts are compared as though they shared a scale, and because congeners differ only on the few attributes that become visible when their descriptions sit side by side.

Across five bird benchmarks and two further domains, marine species and Pokémon, CoLAG consistently outperforms strong baselines, improving over state-of-the-art MLLMs by up to 26.5%.

Our contributions are as follows:

- We systematically characterize the performance degradation of existing VLMs when moving from closed-set to open-vocabulary species recognition, showing that fine-grained distractor confusion is a primary bottleneck.
- We construct a knowledge base of 11,202 bird species with distilled, visually discriminative summaries, validated by human evaluation.
- We propose CoLAG, a two-stage framework combining neighborhood-whitened multimodal retrieval with structured comparative reasoning, recasting recognition as a decision problem over retrieved candidates.
- We demonstrate consistent improvements across seven benchmarks and three domains.

2 Related Work

Open-Vocabulary Recognition. Early work Zhao et al. (2017) introduced joint image-word embeddings, later superseded by large-scale vision-language models such as CLIP Radford et al. (2021), OpenCLIP Ilharco et al. (2021), and SigLIP Zhai et al. (2023), extended to detection Gu et al. (2022), segmentation Li et al. (2022), and classification Dao et al. (2023); Zhu et al. (2024). Aligning images and text at the level of holistic caption-image pairs gives strong zero-shot transfer but lacks the attribute-level discriminative capacity needed when thousands of similar candidates compete. CoLAG combines an ensemble of complementary VLMs for retrieval with a reasoning stage that explicitly compares fine-grained attributes across candidates.

Species Recognition. Benchmarks such as CUB-200-2011 Wah et al. (2011), Birdsnap Berg et al. (2014), and iNaturalist Van Horn et al. (2018) have driven progress in fine-grained classification, yet

evaluations remain predominantly closed-set Radford et al. (2021); Elhoseiny et al. (2013); Naeem et al. (2023). Methods augmenting VLMs with LLM-generated descriptions Menon & Vondrick (2023) or retrieved knowledge Liu et al. (2023) improve within-dataset accuracy but are restricted to predefined class lists, whereas real-world recognition is open-vocabulary: species may be unseen, reclassified, or novel. We evaluate under a unified open-vocabulary protocol in which every query is compared against all 11,202 species, regardless of dataset membership.

MultiModal Large Language Models(MLLM). MLLMs OpenAI (2024a;b); Team et al. (2025a); DeepSeek-AI et al. (2025); Team (2025); Jiang et al. (2024) show strong visual reasoning, but two limitations constrain their use here. They rely on parametric knowledge to recall species facts, which is unreliable for the ever-expanding long tail of visually similar taxa, and their performance degrades with context length Yin et al. (2024), making exhaustive candidate descriptions impractical. CoLAG addresses both: CoLAG$_{ret}$ narrows candidates to a small, high-quality set, and CoLAG$_{gen}$ supplies concise summaries that fit comparative reasoning within the effective context window.

Retrieval Augmented Generation(RAG). RAG frameworks Lewis et al. (2020); Shi et al. (2024) ground generation in retrieved knowledge, with later work improving fusion Izacard & Grave (2021) and extending to multi-hop QA and hallucination mitigation Mallen et al. (2023); Asai et al. (2023); multimodal variants such as MIRAGE Wu et al. (2025) and UNI-RAG Sharifymoghaddam et al. (2025) add image retrieval. Most retrieve independent text chunks and evaluate them in isolation, which suffices for factual recall but not for fine-grained recognition, where the signal lies in relative differences between competing candidates. CoLAG instead retrieves with fused multimodal representations and presents the MLLM with complete summaries for all candidates simultaneously, enabling joint cross-candidate comparison.

3 Dataset and Benchmark

Our benchmark comprises 11,202 species, each paired with a Wikipedia article, a concise summary, and representative images, referred to as anchor images.

Data Collection. We compile all bird species listed on Wikipedia [1] and retrieve the textual and visual content of each page; the 48 species without a dedicated page are excluded. This yields $11,202$ species, each with its Wikipedia description and images, forming the knowledge base for open-vocabulary recognition.

Summary Generation and Visual Refinement. Wikipedia articles carry extraneous information, such as name origins and historical context, that adds context length without visual cues and can therefore hurt performance. We use GPT-4o OpenAI (2024a) to rewrite each article into a concise summary of visually distinguishing attributes (prompt in Section A), reducing the average word count per species from 552 to 127.

The images are equally heterogeneous, including habitat maps, conservation park logos, and scientific illustrations alongside photographs. We passed all 84,993 images to GPT-4o with the prompt "Is there a bird present in the Image? Reply only with Yes or No", retaining only those answered 'Yes'. Filtering removed 49,038 images and left an average of 3 anchor images per species.

Pokémon Dataset. We start from the canonical species list of Pokémondb and generate concise visual summaries with GPT-4o as for birds. Images were scraped from Google Search and Kaggle datasets, but often contained multiple Pokémon or the wrong species, so all were manually verified against the representative image on the Pokémondb Pokedex page and inaccurate samples discarded. The result is a test set of 11,216 images over 248 species, providing a controlled setting to validate generalization across a different visual distribution.

Description Quality. We assess description quality with two human protocols. For **helpfulness**, annotators rated descriptions on a 1-5 scale for identifying species across 10% of the bird data and

[1] https://en.wikipedia.org/wiki/List_of_birds_by_common_name

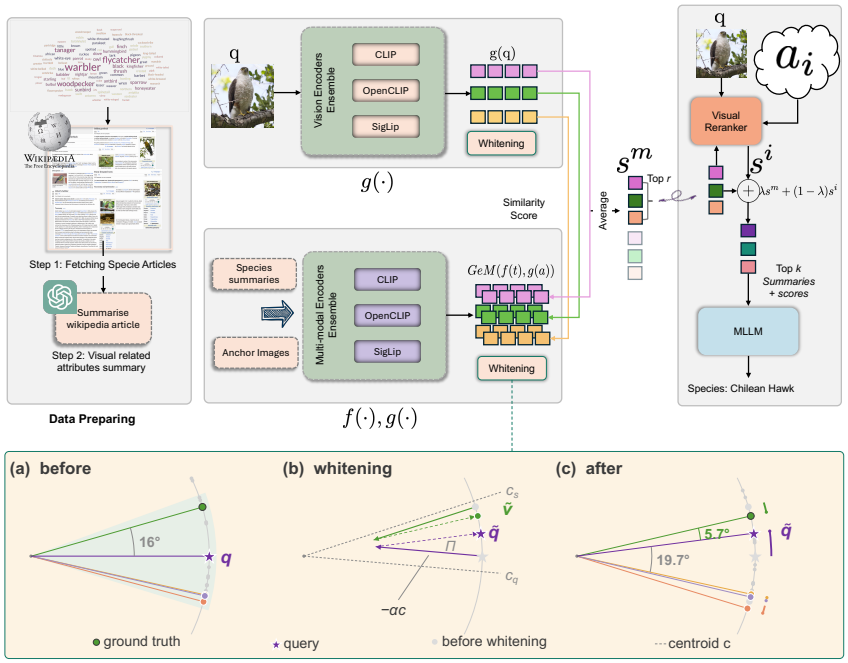


Figure 1: **The CoLAG pipeline**. *Top*: textual and visual extraction from Wikipedia articles (left); similarity to the query from an ensemble of multi-modal encoders with neighborhood whitening (middle); anchor-image re-ranking and MLLM prediction (right). *Bottom*: the WHITENING blocks (Equation (7)) shows (a) the candidates crowd one narrow wedge, so their cosines nearly tie. **(b)** Each item, the query included, has *its own* neighborhood centroid c subtracted and is renormalized. **(c)** Arcs outside the circle show how far each moved: the dense clump barely moves, the query is pushed clear of it.

all 248 Pokémon: birds scored 4.3 on average (median 4.0) and Pokémon 4.5 (median 5.0), with Quadratic Weighted Agreement of 90.6% and 92.2%. For **discriminative matching**, users matched a description to one of five images (one target, four CLIP-space distractors) for 200 CUB species, reaching 94% accuracy at roughly one minute per trial; only 2% were incorrect and 4% failed to converge. We validate the descriptions quantitatively in Section 5.4, with further analysis and qualitative samples in Section K and Section L.

4 METHOD

We introduce a training-free framework, CoLAG(fig. 1), for open-vocabulary species recognition, comprising CoLAG$_{ret}$, which retrieves relevant species from a large database, and CoLAG$_{gen}$, which reasons over them to produce the final answer.

4.1 CoLAG$_{ret}$

Notation. Each species $s \in \{1, \dots, S\}$ consists of a textual description together with anchor images x_s depicting it; a *text-visual chunk* pairs one description t_i with the anchor images of its species. The database holds N such chunks, with $\sigma : \{1, \dots, N\} \to \{1, \dots, S\}$ mapping a chunk to its species ($N = 73{,}051$, $S = 11{,}202$; ≈ 6.5 chunks per species), so retrieval operates on chunks while prediction is over species. Let $\mathcal{M} = \{\text{CLIP}, \text{OPENCLIP}, \text{SIGLIP}\}$ with $M = |\mathcal{M}|$, let $f^{(m)}, g^{(m)}$ be the text and image towers of backbone m, and let $\Pi(x) = x/\|x\|_2$. Fusing the two modalities of a chunk with the signed generalised mean (GeM) gives

$$v_i^{(m)} = \Pi\Big(\text{GeM}\big(f^{(m)}(t_i),\, g^{(m)}(x_{\sigma(i)})\big)\Big) \in \mathbb{S}^{d_m - 1}, \qquad q^{(m)} = \Pi\big(g^{(m)}(q)\big), \tag{1}$$

the GeM-fused, ℓ_2-normalised embedding of chunk i and that of the query q, which is visual only and so passes through the image tower alone. The ensemble retrieval score and the DINO reranking

scores are

$$f(i,q) = \frac{1}{M} \sum_{m \in \mathcal{M}} \langle q^{(m)}, v_i^{(m)} \rangle, \qquad d(i,q) = \langle u_q,\, a_{\sigma(i)} \rangle, \tag{2}$$

where u_q is the DINO query embedding and a_s the average DINO visual anchor of species s, obtained from x_s. With $\mathcal{W}(q)$ the top-K candidates under f ($K = 100$), the reranked score is

$$r(i,q) = \lambda f(i,q) + (1-\lambda)\, d(i,q), \qquad i \in \mathcal{W}(q). \tag{3}$$

We find that the top-5 candidate distribution shows that failures are overwhelmingly *local*: on CUB, the rank-1 prediction is a congener of the ground truth in 44% of failures and a confamilial in 83%, and the correct family appears somewhere in the top-5 in 90% of them. Retrieval reaches the correct region of the taxonomy but cannot discriminate within it. Two distinct geometric pathologies produce this behavior, and we correct each in the space where it lives.

4.1.1 HUBNESS PENALTY IN THE VISUAL SPACE

High-dimensional nearest-neighbor retrieval suffers from *hubness* Radovanovic; et al. (2010): a small number of points are unusually close to many others and therefore appear among the top candidates for queries they are unrelated to. Let $\mathcal{N}_k^{\text{DINO}}(s)$ be the k nearest species to s under $\langle a_s, a_t \rangle$, excluding s itself, and define the *local density*

$$\rho_s = \frac{1}{k} \sum_{t \in \mathcal{N}_k^{\text{DINO}}(s)} \langle a_s, a_t \rangle. \tag{4}$$

We penalize candidates in proportion to how crowded their neighborhood is:

$$h(i,q) = -\rho_{\sigma(i)}. \tag{5}$$

Intuitively, ρ_s measures how much of species s's similarity to a query is explained by s being close to *everything* rather than close to *this* query; subtracting it removes that unearned component. Self-exclusion in $\mathcal{N}_k$ is essential, since $\langle a_s, a_s \rangle = 1$ would otherwise dominate the average. Note that h depends on i only through its species $\sigma(i)$, so the penalty is shared by all chunks of a species and is computed once per species offline.

4.1.2 SYMMETRIC NEIGHBORHOOD WHITENING IN THE LANGUAGE-VISION SPACE

Per-backbone species anchors and their neighborhood centroids are defined as:

$$A_s^{(m)} = \Pi\Big(\textstyle\sum_{i:\sigma(i)=s} v_i^{(m)} \Big), \qquad c_s^{(m)} = \Pi\Big(\frac{1}{k} \textstyle\sum_{t \in \mathcal{N}_k(s)} A_t^{(m)} \Big), \tag{6}$$

where $\mathcal{N}_k(s)$ is now taken in backbone m's own space. Anchors are formed at the *species* level rather than the chunk level because the several chunks of one species describe the same entity; the discriminative question is which species, not which sentence. We then subtract each item's neighborhood centroid from *both* sides of the comparison,

$$\tilde{v}_i^{(m)} = \Pi\big(v_i^{(m)} - \alpha\, c_{\sigma(i)}^{(m)}\big), \qquad \tilde{q}^{(m)} = \Pi\big(q^{(m)} - \alpha\, c_q^{(m)}\big), \tag{7}$$

with $c_q^{(m)} = \Pi\big(\frac{1}{k} \sum_{t \in \mathcal{N}_k(q)} A_t^{(m)}\big)$, and score

$$w(i,q) = \frac{1}{M} \sum_{m \in \mathcal{M}} \langle \tilde{q}^{(m)},\, \tilde{v}_i^{(m)} \rangle. \tag{8}$$

Equation equation 7 is an unsharp mask on the embedding manifold: it removes the component each item shares with its neighborhood and keeps what distinguishes it, thereby increasing discriminatory features. The symmetry in equation 7 is not cosmetic. Correcting only the database cannot reorder the confusable pairs that the method relies on. Empirically, one-sided whitening harms the performance on every dataset (Table 12), while the symmetric form improves with a mean gap of $+0.0284$ hit@1 between the two variants. The proof for this phenomenon is given in Section F.

4.1.3 COMBINATION

The two corrections have different units (h is a density, w a cosine), so we standardise each within the candidate window before mixing. Writing $\mu_q[g]$ and $s_q[g]$ for the mean and standard deviation of $g(\cdot, q)$ over $i \in \mathcal{W}(q)$ and $z_q[g](i) = \big(g(i,q) - \mu_q[g]\big)/s_q[g]$, the final score is

$$\boxed{\hat{r}(i,q) = r(i,q) + \beta\, s_q[r]\,\Big(z_q[h](i) + z_q[w](i)\Big)} \tag{9}$$

Two properties make equation 9 convenient. (i) $\beta = 0$ recovers equation 3 exactly, so the baseline is a strict special case. (ii) The factor $s_q[r]$ places the correction on the scale of the existing score, so β is dimensionless and directly interpretable as "fraction of the base score's spread". Only candidates already retrieved are rescored, so equation 9 leaves recall unchanged and adds no risk of degrading the retrieval stage.

4.2 CoLAG$_{gen}$

Standard MLLMs rely on implicit parametric knowledge and struggle to compare fine-grained attributes among visually similar species, causing confusion in open-vocabulary settings. $CoLAG_{gen}$ instead presents the query alongside a ranked set of hypotheses retrieved by $CoLAG_{ret}$. Each candidate includes its species name, a concise visual summary, and a retrieval confidence score. The MLLM compares the image with all descriptions jointly, using retrieval scores as soft priors while overriding them when supported by visual evidence.

This reformulates open-ended generation as comparative reasoning: rather than recalling a species from memory, the model contrasts attributes such as beak shape, plumage pattern, and body proportion. Joint presentation reveals relative differences, for example, shared breast coloring but distinct wing-bar patterns that isolated evaluation may miss. We evaluate the summaries in Section 5.4.

5 EXPERIMENTS AND RESULTS

Below, we outline our experimental setup, detailing the datasets used, evaluation metrics, and baseline methods. We then present our results, accompanied by an ablation study, to demonstrate the effectiveness of each component.

5.1 EXPERIMENTAL SETUP

Datasets. We evaluate CoLAG on five bird benchmarks spanning a range of geographic regions and taxonomic diversity. **CUB-200-2011** Wah et al. (2011) contains 5,794 test images over 200 North and South American species; **Birdsnap** Berg et al. (2014) covers 500 North American species with 2,443 test images; **Indian Birds** Rokde (2023) comprises 7,499 test images over 25 species; **iNaturalist** Van Horn et al. (2018) contributes its Aves class, 13,230 validation images over 1,323 species; and **NaBirds** Van Horn et al. (2015) covers 400 North American species (555 visual categories) with 24,633 test images. To test generalization beyond birds we add **FishNet** Khan et al. (2023), a marine dataset with 17,393 candidate descriptions, and a curated **Pokémon** set of 11,216 images over 248 species with 1,024 candidate descriptions. At test time, every benchmark is evaluated open-vocabulary: each query is ranked against the full knowledge base (11,202 species for birds, 17,393 for FishNet, 1,024 for Pokémon), regardless of whether the species appeared in the original dataset.

Metrics. For retrieval, we report recall@k (k $\in$ 1, 5, 10), the percentage of queries with a correct species in the top-k; for recognition, we report top-1 accuracy.

Baselines. We compare against two families. For retrieval, four VLMs: CLIP Radford et al. (2021), OpenCLIP Ilharco et al. (2021), and SigLIP Zhai et al. (2023) on ViT-L14, and BioCLIP Stevens et al. (2024) on ViT-B16. For recognition, six open-source MLLMs: Gemma-3n-E4B-it Team et al. (2025b), Qwen2.5-VL-7B-Instruct Qwen et al. (2025), InternVL3_5-8B Wang et al. (2025), Molmo2-8B Clark et al. (2026), MiniCPM-V-4_5 Yu et al. (2025), and Qwen3-VL-8B-Instruct Team (2025), each assessed in four settings: (i) direct prediction from the query image

Table 2: **Retrieval Results**: We compare CoLAG$_{ret}$ with cross-modal(text-to-image) VLM retrieval methods across five benchmarks. Results using our multi-modal ℓ_2-normalised embeddings are reported in section I.

	Birdsnap			**CUB**			**iNaturalist Birds**		
	recall@1	recall@5	recall@10	recall@1	recall@5	recall@10	recall@1	recall@5	recall@10
BioCLIP	19.4	39.7	48.8	24.3	43.3	52.4	13.5	29.4	37.9
CLIP	17.2	40.7	51.1	16.1	39.8	51.4	7.7	21.5	29.8
OpenCLIP	15.7	41.6	52.8	15.2	40.5	52.3	8.5	24.2	34.2
SigLIP	18.8	44.6	55.6	20.1	47.1	58.7	11.3	28.6	37.7
CoLAG$_{ret}$	**51.8**	**75.9**	**82.3**	**60.8**	**81.5**	**86.6**	**36.6**	**60.7**	**67.9**
	Indian Birds			**NABirds**			**Average**		
	recall@1	recall@5	recall@10	recall@1	recall@5	recall@10	recall@1	recall@5	recall@10
BioCLIP	23.0	39.6	45.9	21.8	41.8	51.0	20.4	38.8	47.2
CLIP	17.6	41.9	52.3	18.2	42.7	54.0	15.4	37.3	47.7
OpenCLIP	10.2	29.9	41.3	15.5	40.6	53.3	13.0	35.4	46.8
SigLIP	21.9	46.5	58.2	20.5	30.0	31.4	18.5	39.4	48.3
CoLAG$_{ret}$	**68.3**	**88.4**	**89.5**	**54.9**	**78.9**	84.5	**54.5**	**77.1**	**82.2**

alone; (ii) SigLIP-augmented, where SigLIP-retrieved candidates are supplied with their retrieval scores; (iii) candidate-wise, where the MLLM scores each retrieved candidate independently; and (iv) CoLAG, our full pipeline supplying candidates with distilled summaries and retrieval scores. We also compare with RAG-based approaches: MIRAGE, React and Visual Description.

5.2 Results

Retrieval. Table 2 presents retrieval results across all five bird benchmarks. CoLAG$_{ret}$ consistently outperforms all baselines at every recall level. On average, it surpasses the strongest single-encoder baseline (SigLIP) by 36.0 points on recall@1, 37.7 on recall@5, and 33.9 on recall@10. The gains are particularly pronounced on datasets with high visual similarity among species: on CUB, recall@1 improves from 20.1 (SigLIP) to 60.8, and on Indian Birds from 21.9 to 68.3. These results confirm that whitened multimodal representations, combined with visual re-ranking, provides substantially stronger retrieval than any individual VLM operating on text-image alignment alone. Importantly, the high recall@5 of 77.1% on average establishes the effective upper bound for downstream recognition, since CoLAG$_{gen}$ reasons over the top-5 retrieved candidates.

Recognition. Table 3 reports recognition results across five bird benchmarks. CoLAG consistently improves all models and datasets, outperforming direct prompting by 40.1 points for Gemma-3n and 27.9 points for Qwen3-VL on average. Gains across different model capacities indicate that performance stems from CoLAG's candidate retrieval and reasoning rather than model scale. Reliable retrieval is essential: naive SigLIP augmentation provides limited gains and can even hurt performance, reducing Qwen2.5-VL from 36.9 to 29.6, because noisy candidates become distractors. CoLAG instead retrieves candidates using both visual and textual evidence, making the correct species more likely to be included.

Candidate quality alone is insufficient; their presentation also matters. CoLAG$_{gen}$ jointly presents all k candidate summaries, whereas the candidate-wise baseline scores them independently and selects the $\arg\max$. Although candidate-wise scoring improves direct prompting, it trails joint reasoning by 11.3 points for Qwen3-VL because scores from separate contexts are not necessarily calibrated. Joint presentation enables direct comparison of discriminative attributes, which is crucial because most errors involve congeners with highly overlapping descriptions (Sec. H).

Together, reliable retrieval and joint comparison allow all open-source models using CoLAG to outperform GPT-5-mini on average, showing that structured retrieval and comparative reasoning can compensate for model scale.

RAG Baselines. Table 4 compares CoLAG against RAG-based approaches. CoLAG reaches 58.4% average accuracy, 27.8 points above the strongest baseline (Visual Description Menon & Vondrick (2023), 30.6%). MIRAGE Wu et al. (2025) performs poorly because its lightweight retriever is not equipped for open-vocabulary fine-grained retrieval. Both retrieval quality and reasoning over structured candidate descriptions are essential.

Table 4: **Comparison with other RAG Methods.**

	Birdsnap	**CUB**	**iNat.**	**Ind. Birds**	**NABirds**	**Average**
MIRAGE Wu et al. (2025)	4.1	4.4	3.4	3.9	4.4	4.0
ReactLiu et al. (2023)	28.9	30.2	16.7	24.9	32.3	26.6
Visual DescriptionMenon & Vondrick (2023)	34.7	36.2	17.6	26.1	38.3	30.6
Qwen3-VL(CoLAG)	**57.2**	**62.1**	**39.6**	**72.7**	**60.4**	**58.4**

Table 3: **Performance comparison with MLLMs**. CoLAG consistently improves across all MLLMs on all five benchmarks.

Models	Size	Birdsnap	CUB	iNat	Indian	Nabirds	Average
			Proprietary MLLM				
GPT-5-mini	-	52.9	54.4	23.9	61.6	59.1	50.4
			Open-Source MLLMs				
Gemma-3n Team et al. (2025b)	4B	17.1	19.7	8.3	13.1	22.9	16.2
Gemma-3n + SigLIP	4B	28.4	31.4	16.7	31.1	32.2	28.0
Gemma-3n + Candidate-wise	4B	49.1	54.7	32.1	57.5	49.9	48.7
Gemma-3n(CoLAG)	4B	**54.0**	**62.6**	**37.8**	**70.3**	**57.0**	**56.3**
Qwen2.5-VL Qwen et al. (2025)	7B	39.2	44.9	17.9	33.1	49.2	36.9
Qwen2.5-VL + SigLIP	7B	30.4	32.8	16.9	34.1	33.9	29.6
Qwen2.5-VL + Candidate-wise	7B	44.6	50.9	30.1	57.6	49.1	46.5
Qwen2.5-VL(CoLAG)	7B	**53.9**	**63.4**	**37.7**	**71.9**	**56.2**	**56.6**
InternVL3_5 Wang et al. (2025)	8B	6.8	8.6	2.2	1.1	9.4	5.6
InternVL3_5 + SigLIP	8B	27.1	29.4	14.6	28.9	29.3	25.9
InternVL3_5 + Candidate-wise	8B	32.0	38.8	21.7	38.7	35.2	33.3
InternVL3_5(CoLAG)	8B	**50.2**	**57.1**	**33.3**	**61.5**	**53.5**	**51.1**
Molmo2 Clark et al. (2026)	8B	10.7	9.1	3.6	0.6	13.7	7.5
Molmo2 + SigLIP	8B	24.6	26.5	14.5	29.5	26.5	24.3
Molmo2 + Candidate-wise	8B	46.6	52.5	30.9	59.3	47.2	47.3
Molmo2(CoLAG)	8B	**50.1**	**57.6**	**35.0**	**68.3**	**53.4**	**52.9**
MiniCPM-V-4_5 Yu et al. (2025)	8B	27.1	28.4	10.7	19.2	33.5	23.8
MiniCPM-V-4_5 + SigLIP	8B	30.8	33.0	17.2	33.3	33.6	29.6
MiniCPM-V-4_5 + Candidate-wise	8B	42.5	47.1	27.5	57.0	43.5	43.5
MiniCPM-V-4_5(CoLAG)	8B	**54.0**	**61.8**	**37.7**	**69.3**	**56.8**	**55.9**
Qwen3-VL Team (2025)	8B	35.2	33.8	14.7	29.1	39.6	30.5
Qwen3-VL + SigLIP	8B	36.4	37.6	20.4	41.1	39.6	35.0
Qwen3-VL + Candidate-wise	8B	46.7	51.3	31.3	57.3	48.9	47.1
Qwen3-VL(CoLAG)	8B	**57.2**	**62.1**	**39.6**	**72.7**	**60.4**	**58.4**

5.3 Additional Domains

Table 5: Classification accuracy for the FishNet and Pokémon datasets.

	FishNet	Pokémon
Qwen3-VL Team (2025)	9.4	48.2
Qwen3-VL + SigLIP	5.9	83.3
Qwen3-VL + Candidate-wise	15.1	82.7
Qwen3-VL(CoLAG)	**19.2**	**85.0**

We report results on FishNet and the Pokémon dataset in Table 5, for the best-performing MLLM(Qwen3-VL). The poor performance on FishNet reflects the lack of marine domain knowledge and the task's extreme fine-grained nature, which spans over $17,000$ species. Despite this, CoLAG improves upon Qwen3-VL by 9.8%. On the Pokémon dataset, while the initial performance is better than FishNet, CoLAG yields a huge boost of 36.8%, demonstrating that CoLAG is effective across diverse domains. Detailed results for all MLLMs are provided in Section B.

5.4 Distilling Wikipedia.

Table 6: Context comparison.

	Birdsnap	CUB	iNat.	Ind. Birds	NABirds	Average
Qwen3-VL(Names)	47.3	47.8	31.5	58.7	51.0	47.3
Qwen3-VL(Chunks)	51.5	54.6	34.0	63.4	52.7	51.2
Qwen3-VL(Raw Wikipedia)	54.2	60.4	38.8	70.0	57.7	56.2
Qwen3-VL(CoLAG)	**57.2**	**62.1**	**39.6**	**72.7**	**60.4**	**58.4**

Table 6 compares four levels of textual context, all using CoLAG$_{ret}$ for retrieval. Species names alone already add 16.8 points over the base model (47.3% vs. 30.5%), confirming the value of narrowing the label space. Retrieved text chunks push this to 51.2%, and full Wikipedia articles to 56.2%, at substantial token cost. Our distilled summaries do best (58.4%) while using roughly 75% fewer tokens per species than raw articles, showing that concise, visually discriminative descriptions trade off informativeness and efficiency better than large, noisy context.

Table 7: Ablation study: We evaluate each of CoLAG module's impact for the recall@1, recall@5, and recall@10 on CUB-200-2011 Wah et al. (2011), Birdsnap Berg et al. (2014), and Indian Birds Rokde (2023). The best results are highlighted in bold text for each column.

CLIP	OpenCLIP	SigLIP	Multimodal	Re-Ranker	Hubess & Whitening	Birdsnap Berg et al. (2014)			CUB Wah et al. (2011)		
						recall@1	recall@5	recall@10	recall@1	recall@5	recall@10
✓	✗	✗	✗	✗	✗	17.2	40.7	51.1	16.1	39.8	51.4
✓	✓	✗	✗	✗	✗	25.5	54.5	65.0	25.5	55.3	65.9
✓	✓	✓	✗	✗	✗	28.8	59.1	68.7	30.0	61.1	72.1
✓	✓	✓	✓	✗	✗	40.9	66.6	74.8	49.7	74.1	81.6
✓	✓	✓	✓	✓	✗	49.5	72.9	80.6	58.9	80.1	85.5
✓	✓	✓	✓	✓	✓	**51.8**	**75.9**	**82.3**	**60.8**	**81.5**	**86.6**

5.5 Analysis of System Upper Bound.

CoLAG's accuracy is bounded by retrieval: if the ground-truth species is absent from the top-k candidates, the MLLM cannot recover it. Since CoLAG$_{gen}$ reasons over the top-5, the average recall@5 of 77.1% (Table 2) is the ceiling, and CoLAG's 58.4% reaches roughly 76% of it. The gap has two sources: retrieval errors, where the correct species never enters the candidate set (22.9% of queries), and reasoning errors, where it is present but not selected. Retrieval is thus the dominant bottleneck, though sharper fine-grained reasoning in MLLMs remains necessary to close the remainder.

5.6 Ablations.

Retrieval Components. Table 7 isolates each component of CoLAG$_{ret}$. Starting from CLIP alone, every addition, OpenCLIP, SigLIP, multimodal embeddings, visual re-ranking, and hubness penalty with whitening, yields consistent incremental gains, distributed across components rather than concentrated in one. MLLM performance under each retrieval configuration is in Section C, and a comparison of vision encoders for re-ranking in Section G.

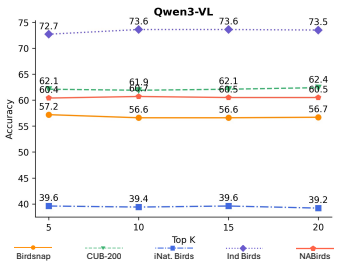


Figure 2: Impact of top-k candidates(5 to 20) fed to Qwen3-VL.

Top-k Selection. Figure 2 shows that Qwen3-VL accuracy is stable for $k = 5$–20, varying by at most 0.9 points. Higher recall from larger candidate sets is offset by additional distractors. We therefore use $k = 5$, which provides comparable accuracy at the lowest inference cost.

Ablating Response Paradigms. Table 8 compares three ways of eliciting the final answer, all given the same top-5 descriptions from CoLAG$_{ret}$. Asking the model to name the species directly beats the multiple-choice formulation, where candidates are labelled A-E and the returned letter is mapped back (56.8 vs. 51.4): indexing by letter adds indirection the model must resolve, and the letters carry no information about the species they stand for. Appending the CoLAG$_{ret}$ retrieval scores gives a further gain (58.4), the scores acting as soft priors that calibrate across closely ranked candidates.

Table 8: Response Paradigms.

	Birdsnap	CUB	iNat.	Ind. Birds	NABirds	Average
Option	50.0	54.2	34.4	65.0	53.3	51.4
Answer	55.4	60.0	38.5	71.6	58.3	56.8
Answer+Scores	**57.2**	**62.1**	**39.6**	**72.7**	**60.4**	**58.4**

6 Conclusion

We present CoLAG, a training-free framework that casts open-vocabulary species recognition as comparative reasoning over a small set of retrieved candidates, grounded in distilled, visually discriminative knowledge. By combining multimodal retrieval with structured cross-candidate comparison, CoLAG consistently outperforms direct MLLM prediction and prior RAG approaches across bird benchmarks, a marine dataset, and Pokémon. Our analysis identifies retrieval as the main bottleneck, with MLLMs achieving ∼76% of the upper-bound accuracy. Future work includes improving retrieval for under-documented taxa and enhancing MLLM reasoning to close this gap.

7 ETHICS

The primary goal of this research is to advance AI for a positive societal impact, specifically in biodiversity conservation. All textual data sourced from encyclopedic resources such as Wikipedia will be made publicly available under an appropriate license. The dataset contains images of animal species and does not involve human subjects, thus presenting no personal data privacy concerns. We believe the potential for misuse of this technology is low, as its primary applications are scientific research and environmental monitoring.

8 REPRODUCIBILITY

We are committed to ensuring reproducibility of our work. All datasets, including the curated Wikipedia-based benchmark and the Pokémon dataset, will be made publicly available under appropriate licenses. We specify hyperparameters and architectures for evaluation in the main text and supplementary material. Our codebase, data loaders, and evaluation scripts will be released on GitHub. Random seeds are fixed in all experiments. Together, these steps ensure that our results can be independently verified and extended by the community.

USE OF LARGE LANGUAGE MODELS

LLMs have been used to summarise the species description, and each use is described in the paper.

- *Dataset cleaning and reformatting:* we used GPT-4o to rewrite the Wikipedia article of each of the 11,202 bird species (and each of the 248 Pokémon species) into a short summary of visually distinguishing features. We also used GPT-4o to filter the 84,993 scraped Wikipedia images with a binary "Is there a bird present in the image?" prompt, which removed 49,038 non-photographic or irrelevant images.
- *Method implementation:* the comparative-reasoning stage of CoLAG uses off-the-shelf multimodal LLMs as inference-time backbones. These models, along with GPT-5-mini, are also evaluated as baselines. No model was fine-tuned.

We did not use generative AI tools to propose research hypotheses, design the methodology or experiments, or interpret results.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023. URL https://arxiv.org/abs/2310.11511.

George F Barrowclough, Joel Cracraft, John Klicka, and Robert M Zink. How many kinds of birds are there and why does it matter? *PLoS One*, 11(11):e0166307, 2016.

Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L. Alexander, David W. Jacobs, and Peter N. Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proc. Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2014.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding, 2026. URL https://arxiv.org/abs/2601.10611.

Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:213–220, 1968. URL https://api.semanticscholar.org/CorpusID:29694079.

Son D. Dao, Dat Huynh, He Zhao, Dinh Phung, and Jianfei Cai. Open-vocabulary multi-label image classification with pretrained vision-language model. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 2135–2140, 2023. doi: 10.1109/ICME55011.2023.00365.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *2013 IEEE International Conference on Computer Vision*, pp. 2584–2591, 2013. doi: 10.1109/ICCV.2013.321.

Alvan R. Feinstein and Domenic V. Cicchetti. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549, 1990. doi: 10.1016/0895-4356(90)90158-l.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=lL3lnMbR4WU.

C. E. Hinchliff, S. A. Smith, J. F. Allman, J. G. Burleigh, R. Chaudhary, L. M. Coghill, K. A. Crandall, J. Deng, B. T. Drew, R. Gazis, K. Gude, D. S. Hibbett, L. A. Katz, H. D. 4th Laughinghouse, E. J. McTavish, P. E. Midford, C. L. Owen, R. H. Ree, J. A. Rees, D. E. Soltis, T. Williams, and K. A. Cranston. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences of the United States of America*, 112(41):12764–12769, Oct 13 2015. doi: 10.1073/pnas.1423041112.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773. If you use this software, please cite it as below.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Lin-*

guistics: Main Volume, pp. 874–880, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.74. URL https://aclanthology.org/2021.eacl-main.74/.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024. URL https://arxiv.org/abs/2401.04088.

Faizan Farooq Khan, Xiang Li, Andrew J. Temple, and Mohamed Elhoseiny. Fishnet: A large-scale dataset and benchmark for fish recognition, detection, and functional trait prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20496–20506, October 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RriDjddCLN.

Haotian Liu, Kilho Son, Jianwei Yang, Ce Liu, Jianfeng Gao, Yong Jae Lee, and Chunyuan Li. Learning customized visual models with retrieval-augmented knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15148–15158, June 2023.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. pp. 9802–9822, 01 2023. doi: 10.18653/v1/2023.acl-long.546.

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=jlAjNL8z5cs.

Muhammad Ferjad Naeem, Muhammad Gul Zain Ali Khan, Yongqin Xian, Muhammad Zeshan Afzal, Didier Stricker, Luc Van Gool, and Federico Tombari. I2MVFormer: Large Language Model Generated Multi-View Document Supervision for Zero-Shot Image Classification . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15169–15179, Los Alamitos, CA, USA, June 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.01456. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.01456.

OpenAI. Gpt-4o: Enhanced multimodal language model. *OpenAI Research*, 2024a. https://openai.com/index/hello-gpt-4o/.

OpenAI. Gpt-4v: Multimodal language model with vision capabilities. *OpenAI Research*, 2024b. https://openai.com/index/gpt-4/.

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 Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

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, 2021. URL https://arxiv.org/abs/2103.00020.

Milos Radovanovic;, Alexandros Nanopoulos, and Mirjana Ivanovic;. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(86):2487–2531, 2010. URL http://jmlr.org/papers/v11/radovanovic10a.html.

Vaibhav Rokde. Indian birds dataset, 2023. URL https://www.kaggle.com/datasets/ichhadhari/indian-birds. Accessed: 2025-01-04.

Sahel Sharifymoghaddam, Shivani Upadhyay, Wenhu Chen, and Jimmy Lin. UniRAG: Universal retrieval augmentation for large vision language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2026–2039, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.108. URL https://aclanthology.org/2025.findings-naacl.108/.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: Retrieval-augmented black-box language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8371–8384, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.463. URL https://aclanthology.org/2024.naacl-long.463/.

Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. BioCLIP: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19412–19424, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, et al. Gemini: A family of highly capable multimodal models, 2025a. URL https://arxiv.org/abs/2312.11805.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian

Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025b. URL https://arxiv.org/abs/2503.19786.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit, 2022. URL https://arxiv.org/abs/2204.07118.

Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 595–604, 2015. doi: 10.1109/CVPR.2015.7298658.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Jul 2011.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E. Gonzalez, Trevor Darrell, and David Chan. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9JCNPFL1f9.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12), November 2024. ISSN 2053-714X. doi: 10.1093/nsr/nwae403. URL http://dx.doi.org/10.1093/nsr/nwae403.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe, 2025. URL https://arxiv.org/abs/2509.18154.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. URL https://arxiv.org/abs/2303.15343.

Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, and Antonio Torralba. Open vocabulary scene parsing. pp. 2021–2029, 10 2017. doi: 10.1109/ICCV.2017.221.

Zhen Zhu, Yiming Gong, and Derek Hoiem. Anytime continual learning for open vocabulary classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

A PROMPTS USED

- For distilling Wikipedia articles into concise, discriminative summaries, we provide GPT-4o with the species Wikipedia article and prompting it with:

> Summarize the following information about the bird species "species name" into a concise paragraph. Focus on the key physical characteristics that distinguish this species from other similar bird species. Highlight features like size, beak shape, plumage color patterns, wing shape, and any other unique traits. The summary should be useful for someone trying to identify the bird species from a photograph.

- For CoLAG$_{gen}$ we use the below prompt:

> Analyze the provided bird image in conjunction with the following k species descriptions. Each species includes a Retrieval Confidence Score (0.0 to 1.0) from an initial visual search system.
> Species Candidates: ¡retrieved candidates descriptions¿
> Instructions: Use the Retrieval Scores as a baseline for likelihood. However, your final decision must be based on a critical comparison of the image features against the descriptions. If the visual evidence contradicts a high-scoring candidate, you must prioritize the visual characteristics. State only the common name of the species depicted in the image, exactly as it appears in the provided descriptions.

B COMPLETE EVALUATION FOR FISHNET AND POKÉMON.

In Table 9, we present the performance of all models on the FishNet and the Pokémon datasets. All models struggle with the highly fine-grained and diverse FishNet dataset. Even in this challenging setting, applying CoLAG significantly improves the performance of MLLMs.

On the Pokémon dataset, all Vision-Language Models except BioCLIP achieve superior performance compared to MLLMs. BioCLIP, having been exclusively trained on natural data from the Tree of Life Stevens et al. (2024); Hinchliff et al. (2015), likely lacks exposure to non-natural domains, which explains its poor performance on this dataset. In contrast, the remaining VLMs perform exceptionally well, outperforming all MLLMs. However, when Qwen2.5-VL is augmented with CoLAG, it achieves the best overall performance, demonstrating the effectiveness of our retrieval-augmented approach even in less-specialized domains compared to VLMs.

Table 9: Complete classification results for the FishNet and the Pokémon dataset.

Models	Size	FishNet	Pokémon
Gemma-3n Team et al. (2025b)	4B	4.7	30.2
Gemma-3n + SigLIP	4B	4.8	76.0
Gemma-3n + Candidate-wise	4B	16.9	79.8
Gemma-3n(CoLAG)	4B	**19.7**	**81.2**
Qwen2.5-VL Qwen et al. (2025)	7B	5.1	29.5
Qwen2.5-VL + SigLIP	7B	4.9	80.3
Qwen2.5-VL + Candidate-wise	7B	13.5	78.7
Qwen2.5-VL(CoLAG)	7B	**19.1**	**85.6**
InternVL3_5 Wang et al. (2025)	8B	2.1	12.5
InternVL3_5 + SigLIP	8B	4.2	74.7
InternVL3_5 + Candidate-wise	8B	12.1	67.5
InternVL3_5(CoLAG)	8B	**16.5**	**80.9**
Molmo2 Clark et al. (2026)	8B	2.0	14.3
Molmo2 + SigLIP	8B	4.2	82.4
Molmo2 + Candidate-wise	8B	17.9	84.9
Molmo2(CoLAG)	8B	**18.1**	**85.5**
MiniCPM-V-4_5 Yu et al. (2025)	8B	5.9	33.4
MiniCPM-V-4_5 + SigLIP	8B	4.9	82.6
MiniCPM-V-4_5 + Candidate-wise	8B	**14.8**	79.2
MiniCPM-V-4_5(CoLAG)	8B	**14.8**	**85.4**
Qwen3-VL Team (2025)	8B	9.4	48.2
Qwen3-VL + SigLIP	8B	5.9	83.3
Qwen3-VL + Candidate-wise	8B	15.1	82.7
Qwen3-VL(CoLAG)	8B	**19.2**	**85.0**

C ABLATION WITH RETRIEVAL COMPONENTS.

In this section, we evaluate the contribution of each component within the entire CoLAG framework using QWEN3-VL. Unlike Table 7 in the main paper, which only analyses the effect of components within the CoLAG$_{ret}$, here we analyze the performance when specific components are removed from the complete CoLAG pipeline. As shown in Table 10, each additional component, OpenCLIP, SigLIP, multimodal fusion, and the re-ranker, contributes to consistent improvements across all five

Table 10: Ablation study: Ablation study on retrieval components in CoLAG with QWEN3-VL. We progressively add components and report accuracy across five benchmarks. Each module contributes to performance gains, with the full pipeline achieving the best results.

CLIP	OpenCLIP	SigLIP	Multimodal	Re-Ranker	Hubness & Whitening	Birdsnap	CUB	iNat	Ind. Birds	NABirds	Average
✓	✗	✗	✗	✗	✗	34.4	35.0	17.0	37.3	38.2	32.4
✓	✓	✗	✗	✗	✗	42.7	43.9	22.2	40.9	46.3	39.2
✓	✓	✓	✗	✗	✗	45.7	46.2	24.8	47.2	48.2	42.4
✓	✓	✓	✓	✗	✗	49.6	51.9	29.3	60.6	53.1	48.9
✓	✓	✓	✓	✓	✗	55.2	59.6	37.7	70.1	58.6	57.0
✓	✓	✓	✓	✓	✓	**57.2**	**62.1**	**39.6**	**72.7**	**60.4**	**58.4**

benchmarks. The complete CoLAG pipeline achieves the highest average performance, demonstrating that all components are integral to its effectiveness.

D Leakage Potential

E Anchor-Image Leakage Analysis

Our knowledge base draws its anchor images from Wikipedia. Some of these may also appear in the test splits of the evaluation benchmarks, which would inflate retrieval accuracy. To measure this, we audited every anchor image against every test image.

Detection protocol. We encode all Wikipedia anchor images with the same DINOv2 ViT-L/14. We then compute the cosine similarity between each anchor and every test image in the five bird benchmarks. Every pair with cosine similarity ≥ 0.8 is checked with two pixel-level tests: perceptual difference hashing (dHash, compared by Hamming distance) and normalized cross-correlation (NCC). A pair counts as a near-duplicate only if it passes both tests [dHash Hamming distance $\leq \mathbf{X}$ and NCC $\geq \mathbf{Y}$].

Of the 18,959 anchor-test pairs with cosine similarity ≥ 0.8, only 33 (0.2%) are true near-duplicates. Table 11 breaks the leaked test images down by dataset. In every benchmark, they make up less than 0.2% of the test set.

Table 11: Near-duplicate anchor images found in each benchmark test set.

Dataset	Test Images	Leaked Images	% of Test Set
Birdsnap	1,851	2	0.11
CUB	5,794	11	0.19
iNaturalist	13,230	1	0.01
Indian Birds	7,499	0	0.00
NABirds	24,633	6	0.02

Removing all leaked test images changes DINOv2 top-1 species retrieval accuracy by only -0.06 points on Birdsnap, -0.07 points on CUB, and 0.00 points on Indian Birds. Near-duplicate matches, therefore, do not explain the re-ranker's gains.

Most high-similarity pairs, including those with cosine similarity between 0.90 and 0.99, are different photographs: different individuals, poses, and backgrounds. We found true duplicates only among pairs whose cosine similarity was close to 1.0.

Each species is represented by the average embedding of all its anchor images (median of 3 per species), not by any single reference image. So even when one anchor matches a test image, it accounts for only part of the species representation, further limiting any leakage effect.

F Proposition

Proposition 1 (Degeneracy of one-sided whitening)**.** *Let species s, t satisfy $c_s = c_t = c$ and $\langle A_s, c\rangle = \langle A_t, c\rangle$. Then for every query q the database-only score $\langle q, \Pi(A_s - \alpha c)\rangle$ preserves the ordering induced by $\langle q, A_s\rangle$.*

Table 12: **Proposition 1, empirically.** Change in accuracy relative to the DINO-rerank baseline of equation 3, each variant evaluated at its own best β. One-sided whitening, the database corrected, the query left untouched, falls below the re-weighting control on every dataset, and both metrics, exactly as the degeneracy argument predicts. Correcting both sides recovers the ordering power and is above the control in all ten cases.

	Δhit@1			Δhit@5		
	one-sided	control	symmetric	one-sided	control	symmetric
CUB	-1.16	$+1.05$	$\mathbf{+2.14}$	-0.05	$+1.28$	$\mathbf{+1.52}$
Birdsnap	-0.97	$+1.40$	$\mathbf{+2.76}$	$+0.43$	$+2.54$	$\mathbf{+3.46}$
iNat	-1.14	$+0.41$	$\mathbf{+1.35}$	-0.54	$+0.77$	$\mathbf{+2.04}$
Indian	$+0.73$	$+1.00$	$\mathbf{+1.55}$	-0.27	$+0.24$	$\mathbf{+0.68}$
NABirds	-0.67	$+1.88$	$\mathbf{+3.20}$	$+0.03$	$+1.99$	$\mathbf{+2.76}$
Mean	-0.64	$+1.15$	$\mathbf{+2.20}$	-0.08	$+1.36$	$\mathbf{+2.09}$

Table 13: **Incremental contribution of each component**, averaged over the five benchmarks. Each row adds one component to the row above. The last column reports what each component achieves when applied to the baseline *alone*. The control adds the β machinery of equation 9 with no transform, and is shown in place rather than as a cumulative step. Every entry is taken at its own best β.

	cumulative			applied alone
	hit@1	hit@5	hit@10	Δhit@1
Ensemble retrieval only	42.06	67.26	75.24	—
+ DINO rerank	51.62	74.43	80.38	—
+ hubness penalty equation 5	52.64	76.08	81.69	$+1.02$
+ symmetric whitening equation 7	**54.54**	**77.28**	**82.43**	$+2.20$

Proof. $\langle q, \Pi(A_s - \alpha c)\rangle = \big(\langle q, A_s\rangle - \alpha\langle q, c\rangle\big)/\|A_s - \alpha c\|$. Since $\|A_s\| = \|A_t\| = 1$ and $\langle A_s, c\rangle = \langle A_t, c\rangle$, we get $\|A_s - \alpha c\| = \|A_t - \alpha c\| =: \kappa$. The scores are then $\kappa^{-1}\langle q, A.\rangle - \kappa^{-1}\alpha\langle q, c\rangle$, an increasing affine map of $\langle q, A.\rangle$ with query-independent coefficients. Order is preserved. □

Mutually confusable congeners are exactly the case $c_s \approx c_t$, so Proposition 1 says one-sided whitening *widens the margin without changing the winner*. Correcting the query breaks the degeneracy: the numerator of $\langle \tilde{q}, \tilde{v}\rangle$ expands to

$$\langle q, v\rangle - \alpha\langle q, c_{\sigma(v)}\rangle \underbrace{- \alpha\langle c_q, v\rangle}_{\text{differs for } v_s \neq v_t} + \alpha^2\langle c_q, c_{\sigma(v)}\rangle, \tag{10}$$

in which the third term depends on v directly rather than through its centroid, and therefore separates s from t even when $c_s = c_t$.

Table 12 tests this prediction. It reports the change in hit@1 and hit@5 relative to the DINO-rerank baseline of equation 3 on all five benchmarks, each variant evaluated at its own best β. The *control* column is the same pipeline with the transform deleted, but the machinery kept: the correction term is the untransformed cosine in the language-vision space, so β can do nothing but re-weight the two signals already present in equation 3. It therefore measures what the correction mechanism buys on its own, and a transform that fails to beat it has not been shown to contribute anything. This is the bar Proposition 1 predicts one-sided whitening cannot clear, since rescaling is all it is capable of.

Table 13 isolates what each component contributes by adding them one at a time. Ensemble retrieval alone reaches 42.06 hit@1 averaged over the five benchmarks; DINO reranking lifts this to 51.62, and the two corrections of equation 9 take it to 54.54. The order in which the components are added is arbitrary. Applied to the baseline alone, the hubness penalty contributes $+1.02$ hit@1 and symmetric whitening $+2.20$; reversing the order gives $+2.20$ for whitening and $+0.72$ for hubness on top of it. Neither decomposition is privileged, which is what it means for the two corrections to act almost independently: each recovers most of its individual gain in the presence of the other.

Table 14: We evaluate different encoders as visual re-rankers and show that DinoV2 outperforms the rest.

Re-ranker	Birdsnap	CUB	iNaturalist	Indian Birds	NABirds	Average
	recall@1/5/10	recall@1/5/10	recall@1/5/10	recall@1/5/10	recall@1/5/10	recall@1/5/10
CLIP	33.0/59.1/68.8	43.9/68.5/76.5	21.1/42.3/51.8	46.0/71.2/77.9	38.4/65.6/75.1	36.5/61.3/70.0
OpenCLIP	35.5/61.1/70.3	46.3/70.6/77.4	22.1/43.5/52.8	46.9/73.7/80.3	40.4/67.0/75.6	38.2/63.2/71.3
SigLIP	37.9/63.3/72.9	47.4/71.8/79.1	23.4/44.5/53.8	48.4/74.7/81.2	43.0/69.1/77.3	40.0/64.7/72.9
Deit-III	26.6/57.7/71.7	32.3/66.0/77.9	17.0/42.2/54.5	41.0/71.0/79.6	29.5/61.7/76.2	29.3/59.7/72.0
DinoV2	**49.5/72.9/80.6**	**58.9/80.1/85.5**	**35.2/58.9/66.6**	**66.6/87.9/89.6**	**53.0/77.0/83.2**	**52.6/75.4/81.1**

G VISION ENCODERS.

In Table 14, we present a performance comparison when different visual encoders are used as re-rankers. We test the VLMs from the CoLAG pipeline, DinoV2 Oquab et al. (2024), and DeiT-III Touvron et al. (2022) as visual encoders. The data clearly shows that DinoV2 significantly outperforms not only DeiT-III but also all the other encoders from various VLMs. This superior performance establishes DinoV2 as a strong candidate for re-ranking, capable of substantially improving the quality of initially retrieved chunks and refining the overall retrieval process.

Table 15: **Results at Genus Level**. We map all predicted species from different models to their respective genera. This helps us understand whether misclassification occurs because the task is difficult or because the models are making unexpected mistakes.

Models	Size	Birdsnap	CUB	iNat	Indian	Nabirds	Average
Qwen3-VL Team (2025)	8B	49.0	44.8	26.4	35.9	52.9	41.8
Qwen3-VL + BioCLIP	8B	72.8	75.3	57.6	81.2	77.8	72.9
Qwen3-VL(CoLAG)	8B	**80.0**	**82.4**	**63.4**	**86.9**	**82.9**	**79.1**

H ANALYZING THE MIS-CLASSIFICATIONS.

In this section, we analyze the misclassifications to uncover interesting patterns. We leverage the taxonomic hierarchy by examining accuracy at the genus level. This approach reveals how often predictions fall within the same genus, as species within a genus often share similar appearances. To calculate genus-level accuracy, we map all predicted species to the genus level and then compute the accuracy. Instead of directly evaluating at the genus level, this approach allows us to analyze behavior during species prediction. The results are presented in Table 15.

As shown in Table 15, mapping predictions to the genus level yields substantial accuracy gains across all models and datasets, indicating that a substantial portion of species-level errors is not arbitrary but rather reflects confusion among closely related species within the same genus. This observation aligns with the inherent difficulty of fine-grained avian classification, in which intra-genus species often share overlapping visual traits such as plumage patterns, morphology, and coloration.

The baseline Qwen3-VL achieves moderate genus-level accuracy (e.g., 49.0 on Birdsnap and 52.9 on Nabirds) but performs notably worse on iNat (26.4), suggesting that its errors on long-tailed, fine-grained datasets are not confined to within-genus confusion but instead reflect more fundamental misidentifications. Augmenting the baseline with BioCLIP improves genus-level accuracy across all benchmarks, demonstrating that stronger visual grounding helps the model remain more closely aligned with the ground truth.

Finally, Qwen3-VL (CoLAG) achieves the highest genus-level accuracy across every dataset. The consistent and substantial margin over both baselines suggests that even when CoLAG misclassifies a species, its predictions tend to remain within the correct genus. This indicates that CoLAG's errors are predominantly fine-grained, taxonomically reasonable confusions rather than unexpected mistakes, a desirable property for fine-grained recognition systems, as it implies that the model has learned to localize the correct taxonomic neighborhood even when exact species identification proves challenging.

Table 16: **Retrieval results.** Species-level recall@$\{1, 5, 10\}$ on five benchmarks. For each backbone, *(text)* embeds each database chunk from its description passage alone, while *(ours)* embeds the text–visual chunk, pairing the passage with the anchor imagery of its species. CoLAG$_{ret}$ ensembles the three general-purpose backbones (CLIP, OpenCLIP, SigLIP) under *(ours)*; BioCLIP is a species-specialised backbone reported for reference and is not part of the ensemble. Best per column in **bold**.

	Birdsnap			**CUB**			**iNaturalist Birds**		
	recall@1	recall@5	recall@10	recall@1	recall@5	recall@10	recall@1	recall@5	recall@10
BioCLIP	19.4	39.7	48.8	24.3	43.3	52.4	13.5	29.4	37.9
BioCLIP(ours)	45.4	72.6	80.6	54.2	78.9	84.6	31.8	57.7	66.1
CLIP	17.2	40.7	51.1	16.1	39.8	51.4	7.7	21.5	29.8
CLIP(ours)	25.9	50.4	60.3	32.1	58.3	67.7	14.5	32.8	42.4
OpenCLIP	15.7	41.6	52.8	15.2	40.5	52.3	8.5	24.2	34.2
OpenCLIP(ours)	33.3	57.8	67.3	40.9	66.5	73.7	18.9	39.5	49.0
SigLIP	18.8	44.6	55.6	20.1	47.1	58.7	11.3	28.6	37.7
SigLIP(ours)	35.1	62.3	72.2	42.2	69.3	79.0	21.6	43.4	53.0
CoLAG$_{ret}$	**51.8**	**75.9**	**82.3**	**60.8**	**81.5**	**86.6**	**36.6**	**60.7**	**67.9**
	Indian Birds			**NABirds**			**Average**		
	recall@1	recall@5	recall@10	recall@1	recall@5	recall@10	recall@1	recall@5	recall@10
BioCLIP	23.0	39.6	45.9	21.8	41.8	51.0	20.4	38.8	47.2
BioCLIP(ours)	60.4	82.4	87.5	50.1	78.2	**85.0**	48.4	73.9	80.8
CLIP	17.6	41.9	52.3	18.2	42.7	54.0	15.4	37.3	47.7
CLIP(ours)	30.9	55.1	63.4	28.8	56.1	66.9	26.4	50.5	60.2
OpenCLIP	10.2	29.9	41.3	15.5	40.6	53.3	13.0	35.4	46.8
OpenCLIP(ours)	35.7	65.0	72.6	35.9	62.5	72.1	32.9	58.3	66.9
SigLIP	21.9	46.5	58.2	20.5	30.0	31.4	18.5	39.4	48.3
SigLIP(ours)	42.5	70.7	78.1	38.3	66.8	76.4	36.0	62.5	71.7
CoLAG$_{ret}$	**68.3**	**88.4**	**89.5**	**54.9**	**78.9**	84.5	**54.5**	**77.1**	**82.2**

I TEXT VS MULTI-MODAL RETRIEVAL

Multi-modal chunks. Pairing each description passage with its species' anchor imagery improves retrieval for every backbone, on every dataset, at every cutoff(table 16). Averaged over the five benchmarks, recall@1 rises by 11.0 points for CLIP, 17.5 for SigLIP, 19.9 for OpenCLIP, and 28.0 for BioCLIP, between $1.7\times$ and $2.5\times$ the text-only score. The effect is not merely additive: with text alone, the four backbones are nearly indistinguishable (13.0-20.4 recall@1), whereas with anchors they span 26.4-48.4. Descriptions of closely related species are worded too similarly for the text channel to separate them, so a backbone's advantage only becomes visible once the appearance the text refers to is present in the same embedding. CoLAG$_{ret}$ attains the best recall@1 on all five datasets and the best average at every cutoff (54.5/77.1/82.2).

Table 17: Results with our GeM-fused, ℓ_2-normalised embeddings.

Models	**Size**	**Birdsnap**	**CUB**	**iNat**	**Indian**	**Nabirds**	**Average**
Qwen3-VL Team (2025)	8B	35.2	33.8	14.7	29.1	39.6	30.5
Qwen3-VL + BioCLIP	8B	51.6	55.5	34.7	67.2	56.2	53.0
Qwen3-VL + CLIP	8B	39.2	43.1	22.6	47.3	44.0	39.2
Qwen3-VL + OpenCLIP	8B	45.1	49.4	25.6	55.5	48.4	44.8
Qwen3-VL + SigLIP	8B	45.1	48.0	27.2	57.2	49.3	45.4
Qwen3-VL(CoLAG)	8B	**57.2**	**62.1**	**39.6**	**72.7**	**60.4**	**58.4**

In table 17, we compare the results of directly prompting Qwen3-VL with using our GeM-fused, ℓ_2-normalised embeddings for the retrieval from various VLMs. The results clearly show that multi-modal embeddings-based retrieval yields huge gains over cross-modal standard retrieval

Acadian Flycatcher



Eastern Wood Pewee: The Eastern Wood Pewee (Contopus virens) is a small North American tyrant flycatcher, measuring 13.5–15 cm in length with a wingspan of 23–26 cm. It is characterized by its gray-olive upperparts and light underparts with an olive wash on the breast. Notable features include two pale wing bars and long primary remiges, giving the wings a slim, pointed appearance. The upper part of the bill is dark, while the lower part is yellowish. This bird is often confused with the Western Wood Pewee, but can be distinguished by its distinct mournful whistled "pee-a'wee" call. Unlike the Eastern Phoebe, which lacks defined wing bars and frequently bobs its tail, the Eastern Wood Pewee has a more pronounced primary projection. The Least Flycatcher, another similar species, has a bold eye ring and shorter primary projection, appearing blunt-winged.

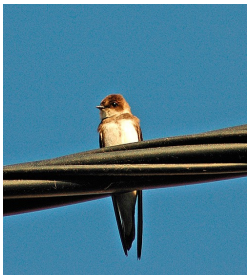
Alder Flycatcher: The Alder Flycatcher (Empidonax alnorum) is a small bird, measuring 13–17 cm in length with a wingspan of 8.3–9.4 cm. It is characterized by its dull greenish-olive upperparts and a slightly darker olive-green crown. The throat is white, contrasted by a darker breast band, and it features thin white eye rings. Its bill is distinctive, with a pinkish or yellow-orange lower mandible and a black upper mandible. The wings are black with white wing bars and white-edged tertials. Juveniles display brownish upperparts and yellow underparts with yellowish-brown or buff wingbars. The Alder Flycatcher is similar to the Eastern Wood-Pewee and Eastern Phoebe but can be distinguished by its smaller size and unique upward tail flicking behavior. Vocalizations, particularly its "fee-bee-o" song, are crucial for identification, as visual differentiation from the nearly identical Willow Flycatcher can be challenging.

Least Flycatcher: The Least Flycatcher (Empidonax minimus) is a small bird, measuring 12 to 14 cm in height with a wingspan of 19 to 22 cm, and weighing around 10.3 grams. It is the smallest of the Empidonax flycatchers in eastern North America. Its plumage is characterized by a dull olive-gray back and a whitish belly, which is notably brighter than other Empidonax species. Distinctive features include two white wing bars and white eye rings. The beak is short, with a yellowish lower mandible. Juveniles resemble adults but have slightly darker wing bars with tawny or olive hues. While these physical traits are shared with other flycatchers, the Least Flycatcher is best identified by its distinctive "che-bec" call and its preferred habitats.

Galapagos flycatcher: The Galápagos flycatcher, or large-billed flycatcher (Myiarchus magnirostris), is a small bird endemic to the Galápagos Islands, measuring 15 to 16 cm in length and weighing between 12 to 18.5 grams. It is the smallest member of its genus. This bird is distinguished by its relatively large bill, which is a key feature for identification. The plumage is generally a muted brown with lighter underparts, and it has a subtle crest on its head. The wings are rounded, typical of flycatchers, aiding in its agile flight. The Galápagos flycatcher is known for its curious behavior, often approaching humans and interacting with camera lenses, mistaking its reflection for another bird.

Western Wood Pewee: The Western Wood Pewee (Contopus sordidulus) is a small tyrant flycatcher, measuring 5.5–6.3 inches in length with a wingspan of about 10.2 inches. It is characterized by its gray-olive upperparts and light underparts, with an olive wash on the breast. Distinctive features include two wing bars and a dark bill with yellow at the base of the lower mandible. This bird closely resembles the Eastern Wood Pewee, but can be distinguished by its call, a loud buzzy "peeer," and its song, which consists of three rapid descending "tsees" ending with a descending "peeer." These features are crucial for visual identification, especially when observing the bird in its natural habitat.

Bank Swallow



Northern Rough Winged Swallow: The Northern Rough-winged Swallow (Stelgidopteryx serripennis) is a small swallow, measuring 13–15 cm in length with a wingspan of 27–30 cm. It is characterized by its brown upperparts and white underparts, with a distinctive brownish-grey wash on the throat. The swallow has a small bill and a forked tail. A key feature is the serrated edge feathers on its wings, which are not visible unless the bird is held. Males have longer and broader undertail coverts compared to females, and their outer primary wings have hooked barbs. Juveniles can be identified by their reddish-brown wing-bars. This species is similar to the Bank Swallow but can be distinguished by its dusky throat and breast. The Northern Rough-winged Swallow is often found near water, where it feeds on flying insects.

Streak-throated swallow: The streak-throated swallow, or Indian cliff swallow (Petrochelidon fluvicola), is a small passerine bird approximately 11 cm (4.3 in) in length, notable for its distinctive physical features. It has a dull chestnut forehead and crown, contrasting with its steel-blue upperparts. The underparts are whitish, and the bird is characterized by brown streaks on the throat and chest, which are key identifiers. Its tail is square-ended, and the rump is pale brown. The streak-throated swallow is often seen in large flocks, especially on cold mornings, perched closely on wires. Its vocalizations include a twittering chirp when perched and a sharp "trr-trr" in flight. These features, along with its habitat preference for open areas near water bodies, help distinguish it from other swallows.

Brown Chested Martin: The brown-chested martin (Progne tapera) is a medium-sized swallow distinguished by its brown chest and upperparts, contrasting with a paler underbelly. It features a streamlined body typical of swallows, with long, pointed wings that are often seen bowed during flight. The bird's tail is slightly forked, displaying white on the sides, which is particularly noticeable when it swoops at low heights. Its beak is short and broad, adapted for catching insects mid-flight. This martin is often observed in large flocks, especially around dusk, and is known for nesting in burrows dug into banks or using old nests of other birds like the hornero. These characteristics, along with its habitat preferences, help distinguish it from other similar species.

Brown-bellied swallow: The Brown-bellied Swallow (Orochelidon murina) is a small bird distinguished by its unique plumage and body shape, which aids in its identification. This species features a distinctive brown belly, contrasting with its darker upperparts, which are typically a shade of blue or black. The swallow's wings are long and pointed, characteristic of the Hirundinidae family, allowing for agile flight. Its beak is short and slightly curved, adapted for catching insects mid-air. The bird's tail is slightly forked, a common trait among swallows, but less pronounced than in some other species. These features, combined with its habitat preference for high-elevation shrublands and grasslands in countries like Bolivia, Colombia, Ecuador, Peru, and Venezuela, make the Brown-bellied Swallow identifiable in its natural environment.

Chilean swallow: The Chilean swallow (Tachycineta leucopyga) is a small bird, approximately 13 centimeters (5.1 inches) in length, with a weight ranging from 15 to 20 grams (0.53 to 0.71 ounces). It is characterized by its glossy blue-black upperparts and white underparts, with a distinctive white rump. The wings and tail are black, featuring white tips on the inner secondaries and tertials, while the underwing coverts and auxiliaries are grey. Both sexes appear similar, though juveniles are duller and browner. Unlike the similar white-rumped swallow, the Chilean swallow lacks a white forehead and has bluer upperparts. Its bill and legs are black, aiding in its identification.

Chestnut Sided Warbler



Reiser's tyrannulet: Reiser's tyrannulet (Phyllomyias reiseri) is a small bird measuring 11 to 11.5 cm in length and weighing 7 to 8 grams. It features a bright olive crown, nape, back, and rump with faint grayish tips on the crown feathers. The bird's face is marked by yellowish white lores, supercilium, and cheeks, with a thin olive line running through the eye. Its wings are dusky with pale yellowish edges on the flight feathers and the ends of the coverts, forming two distinct bars on the closed wing. The tail is dusky olive, while the throat and lower face are whitish, and the underparts are pale yellow with faint olive streaks on the breast and sides. The short, rounded bill has a blackish maxilla and a black-tipped pinkish to white mandible, and the legs and feet are gray. These features, along with its pale brown iris, help distinguish Reiser's tyrannulet from other similar species.

Yellow Throated Vireo: The Yellow-throated Vireo (Vireo flavifrons) is a small songbird distinguished by its striking coloration and features. It measures between 5.1 to 5.9 inches in length with a wingspan of about 9.1 inches. This bird is primarily olive on the head and upperparts, contrasted by a vivid yellow throat and white belly. One of its most distinctive features is the yellow "spectacles" around its dark eyes. The tail and wings are dark, accented with white wing bars, and it possesses thick blue-grey legs and a stout bill. These characteristics, particularly the yellow throat and spectacles, make it easily identifiable among similar species.

Pine Warbler: The Pine Warbler (Setophaga pinus) is a small songbird distinguished by its white belly, two white wing bars, and dark legs. It features a thin, relatively long pointed bill and yellowish "spectacles" around its eyes. Adult males are notable for their olive upperparts and bright yellow throats and breasts, while females and immatures have olive-brown upperparts with paler throats and breasts. This bird's appearance can be confused with the yellow-throated vireo due to similar coloration. The Pine Warbler's song is a musical trill, and its calls are slurred chips. These birds are typically found in open pine woods and are known for their slow foraging on tree trunks and branches, often poking their bills into pine cones.

White Eyed Vireo: The White-eyed Vireo (Vireo griseus) is a small songbird, measuring 4.3–5.1 inches in length with a wingspan of 6.7 inches. It is distinguished by its greyish olive head and back, white underparts with yellow flanks, and dark wings and tail. Notably, it features two white wing bars on each wing and striking white irises surrounded by yellow spectacles, which are key identifiers. Both sexes appear similar. This bird's compact size, distinctive eye markings, and color pattern make it recognizable among other vireos.

Providencia vireo: The Providencia vireo (Vireo approximans) is a small bird distinguished by its robust, thick bill, which is a key feature for identification. It has a predominantly olive-green plumage on its upperparts, with a contrasting pale yellow or whitish underbelly. The wings are short and rounded, typical of vireos, and often display faint wing bars. Its head is marked by a subtle, yet distinctive, white or pale eyebrow stripe above the eye, which helps differentiate it from similar species. The combination of its thick bill, olive-green and yellowish coloration, and the subtle facial markings are essential for identifying the Providencia vireo, especially in its natural habitats of subtropical or tropical moist forests on Isla de Providencia.

Figure 3: Examples of top-5 retrieved candidates that do not contain the correct species description. The retrieved images and their corresponding text descriptions show significant visual and textual similarities to the query image, highlighting the fine-grained and challenging nature of the retrieval task.

Bewick Wren



Marsh Wren: The Marsh Wren (Cistothorus palustris) is a small songbird native to North America, easily identifiable by its distinctive physical features. It measures between 3.9 to 5.5 inches in length with a wingspan of approximately 5.9 inches. This bird has brown upperparts, a light brown belly and flanks, and a white throat and breast. Its back is marked with black and white stripes, and it sports a dark cap with a prominent white line above the eyes. The Marsh Wren has a short, thin bill, which is a key feature distinguishing it from similar species. Its habitat typically includes marshes with tall vegetation like cattails, where it is often seen foraging close to water.

Bewick Wren: The Bewick's Wren (Thryomanes bewickii) is a small North American bird, measuring about 14 cm (5.5 inches) in length with a wingspan of 18 cm. It is characterized by its grey-brown upperparts and white underparts, with a distinctive long white eyebrow. Its long tail, which is frequently moved and features black bars with white corners, is a key identifying feature. The bird has a long, slender, slightly curved beak. Juveniles resemble adults but have shorter, stockier beaks and may show faint speckling on the underbelly. Both males and females look similar. The Bewick's Wren is often confused with the Carolina Wren but can be distinguished by its longer tail and the white tips. It is a vocal species, with males known for their varied and melodious songs.

Carolina Wren: The Carolina Wren (Thryothorus ludovicianus) is a relatively large wren, measuring 12.5 to 14 cm in length with a wingspan of about 29 cm. It is characterized by its rich brown plumage, which appears more chestnut on the rump and upper tail coverts, and a distinctive white superciliary streak bordered by thin black lines. The bird's underparts are a warm buff color, transitioning from gray on the chin and throat. Its wings and tail feature darker brown barring, and the tail has 18 to 20 bars. The Carolina Wren has a slightly curved, lemon-colored upper mandible, and flesh-colored legs. Males are generally larger than females, with longer wing chords. The species can be confused with Bewick's Wren, which is smaller with a longer tail and grayer-brown above. The Carolina Wren's energetic behavior, distinctive song, and preference for dense cover in forests and suburban areas further aid in its identification.

Saint Lucia wren: The Saint Lucia wren (Troglodytes mesoleucus) is a very small bird native to the Caribbean island of Saint Lucia. It is part of the wren family, Troglodytidae, and is distinguished by its diminutive size and tendency to disappear into crevices, a behavior linked to its name meaning "hole dweller." This species features a plumage with a distinctive "middling-white" or streaked white pattern, which sets it apart from other wrens. The Saint Lucia wren has a compact body, a short tail, and a slender, slightly curved beak, typical of wrens. Its wings are short and rounded, aiding in its agile movements through dense vegetation. These physical characteristics, along with its unique vocalizations and behaviors, help in identifying the Saint Lucia wren from photographs.

Kalinago wren: The Kalinago wren (Troglodytes martinicensis) is a very small bird native to the Caribbean island of Dominica, previously also found on Martinique and Guadeloupe. This wren is distinguished by its compact size and characteristic behavior of disappearing into crevices. It features a short, slightly curved beak ideal for insect hunting. The plumage is generally brown with subtle variations, providing excellent camouflage against its natural habitat. The wings are short and rounded, typical of wrens, aiding in quick, agile movements through dense foliage. The Kalinago wren's distinct vocalizations and genetic differences set it apart from the closely related northern house wren.

Baltimore Oriole



Baltimore Oriole: The Baltimore Oriole (Icterus galbula) is a medium-sized passerine bird, measuring 17–22 cm in length with a wingspan of 23–32 cm. It is distinguished by its striking plumage, particularly in males, which feature vibrant orange underparts, shoulder patches, and rump, contrasted with black on the rest of the body. Females and juveniles are more subdued, with yellow-brown upper parts and dull orange-yellow on the breast and belly. Both sexes have distinctive white wing bars. The oriole's thick, pointed bill is typical of the icterid family. This bird is often found high in deciduous trees and is known for its sweet, whistling song. The Baltimore Oriole's preference for ripe, dark-colored fruits and its unique feeding behavior, such as "gaping" into soft fruits, further aid in its identification.

Bullocks Oriole: Bullock's Oriole (Icterus bullockii) is a small, sexually dimorphic bird distinguished by its striking plumage and size. Adult males are notable for their vibrant orange and black coloration, featuring a black throat patch, a black line extending from each eye to the crown, and a white wing bar. Their underparts, breast, and face are orange or yellow, contrasting with the black back, wings, and tail. The tail is mostly black with the outermost rectrices tipped in orange, forming a "T" shape. Females, in contrast, have gray-brown upperparts, duller yellow on the breast and underparts, and an olive crown, lacking the black eye-line seen in males. Both sexes have a pointed bill with a straight culmen, and the wingspan measures about 12.2 inches. Juveniles resemble adult females but have darker wings and a pink or whitish bill. These orioles are often found in riparian corridors and open woodlands, and they are known for their distinctive song, which is similar to but faster and harsher than that of the Baltimore Oriole.

Scott Oriole: The Scott's Oriole (Icterus parisorum) is a medium-sized bird distinguished by its striking plumage and distinct physical features. It measures approximately 9.1 inches (23 cm) in length, with a wingspan of about 12.6 inches (32 cm), and weighs between 1.1 to 1.4 ounces (32–41 g). This oriole is notable for its vibrant yellow and black coloration; the males typically exhibit a bright yellow body with contrasting black head, throat, and upper back, while females and juveniles display a more subdued yellow with olive tones. The beak is slender and slightly curved, typical of the oriole family, aiding in its identification. The wings are pointed, enhancing its agile flight. These distinctive features make the Scott's Oriole easily recognizable, especially in its preferred habitats in the Southwestern United States and parts of Mexico.

Orange oriole: The Orange Oriole (Icterus auratus) is a small bird, measuring about 19–21 cm (7.5–8.3 in) in length, distinguished by its vibrant orange plumage and black markings. The males feature a striking black region between their eyes and bill, with a black stripe across the upper breast, while the head and body are predominantly orange. The wings display a unique pattern with orange lesser feathers and white medium and greater layers. Females are generally duller, with an orange-yellow wash and green mantles. The bird has a slender body, long wings, and a pointed beak, which it uses to pry into fruits. The Orange Oriole's habitat in the Yucatán Peninsula influences its plumage color due to dietary variations. This species is non-migratory, residing in forests, woodlands, and abandoned farmlands.

Altamira Oriole: The Altamira oriole (Icterus gularis) is the largest oriole in its genus, measuring about 25 cm (9.8 in) in length. Both males and females share similar striking features, including a black mandible, throat, back, and long tail. Their wings are predominantly black with white-fringed flight feathers, forming a distinctive single white wing bar and white wing spots when folded. The secondary coverts display bright orange epaulets, while the underside is a vibrant orange or yellowish-orange. Immature birds have an olive back and a dull yellow head and body, with first-year birds resembling adults but with an olive back and yellow-olive tail. These features, along with their large size and vivid coloration, make the Altamira oriole easily distinguishable from other orioles.

Blue Jay



Blue Jay: The Blue Jay (Cyanocitta cristata) is a striking bird native to eastern North America, easily identifiable by its vibrant plumage and distinctive features. Measuring 22–30 cm in length with a wingspan of 34–43 cm, it showcases a prominent blue crest on its head, which can be raised or lowered depending on its mood. The bird's plumage is a captivating lavender-blue to mid-blue on the crest, back, wings, and tail, contrasted by a white face and underparts. A notable black, U-shaped collar encircles its neck, extending to the sides of the head. The wings and tail are intricately barred with black, sky-blue, and white patterns. Its bill, legs, and eyes are black, and both sexes appear similar, with males being slightly larger. The Blue Jay's coloration is due to structural coloration rather than pigments, as the blue hue results from light interference within the feather structure. This bird is known for its noisy and bold behavior, often mimicking the calls of hawks and other sounds.

Eurasian Jay: The Eurasian jay (Garrulus glandarius) is a medium-sized bird in the crow family, measuring 34–35 cm in length with a wingspan of 52–58 cm. It is distinguished by its pinkish-brown body plumage, a whitish throat bordered by black moustache stripes, and a striking bright blue patch on the upper wing with fine black bars. The bird's tail is predominantly black, and its forehead and crown feature whitish feathers with black streaks. The Eurasian jay's complex wing coloration includes black and white bars, and its rump is white. This bird is known for its harsh, rasping screech and its ability to mimic other species' calls. It inhabits mixed woodlands, particularly those with oaks, and is a habitual acorn hoarder. The Eurasian jay's distinct racial forms vary across its vast range, with notable differences in crown streaking, mantle color, and facial markings.

Black-headed Jay: The Black-headed Jay (Garrulus lanceolatus) is a slender bird similar in size to the Eurasian Jay but distinguished by its shorter, thicker bill and more pronounced crest. Its most notable feature is the black top of its head, which contrasts with its otherwise lighter plumage. The bird also has a longer tail compared to its relatives. It inhabits regions from eastern Afghanistan across the Himalayas to Nepal and Bhutan, favoring wooded areas with open spaces. The Black-headed Jay's vocalizations are akin to the Eurasian Jay's, characterized by loud screeches with longer pauses. These features, along with its habitat preferences, make it identifiable from photographs.

White-throated Magpie-Jay: The White-throated Magpie-Jay (Cyanocorax formosus) is a strikingly large bird, measuring between 43 and 56 cm in length, with a distinctive long tail and a slightly curved crest of feathers on its head. The crest is black in the nominate subspecies, with blue or white margins in others. It features a white face contrasted by a black crown and a narrow black band around the throat, with a small black drop below the eye. The breast, belly, and underside of the rump are white, while the wings, mantle, and tail are a vibrant blue, with whitish margins on the tail. The bird's legs and eyes are black, and it has a grey bill. Females resemble males but have duller plumage on top, a narrower chest band, and a shorter tail. These features, along with its gregarious nature and noisy behavior, make the White-throated Magpie-Jay easily identifiable.

Steller's Jay: Steller's jay (Cyanocitta stelleri) is a striking bird native to western North America, easily identifiable by its unique physical characteristics. It measures about 30–34 cm (12–13 in) in length and weighs between 100–140 g (3.5–4.9 oz). This species is the only crested jay west of the Rocky Mountains, featuring a prominent crest that is more pronounced in northern populations. The head varies in color from blackish-brown to dark blue, depending on the subspecies, with lighter streaks on the forehead. The body transitions from a dark head to a silvery blue on the shoulders and lower breast, while the primaries and tail are a rich blue with darker barring. Steller's jays in the eastern part of their range may have white markings on the head, particularly over the eyes, whereas those further west have light blue markers or none at all. The bird's slender bill and longer legs, along with its vibrant plumage and distinctive vocalizations, make it distinguishable from the closely related blue jay.

Figure 4: Examples of top-5 retrieved candidates that contain the correct species description ranked. The other four candidates belong to the same genus as the query image and exhibit significant visual and textual similarities to it, highlighting the fine-grained and challenging nature of the retrieval task.

J THE RE-RANKING MODULE

The re-ranking module begins by selecting the top-r candidate species. For each of these species, the intra-modal similarity is calculated between the anchor images and the query image. A visual illustration of this process is provided in Figure 5.

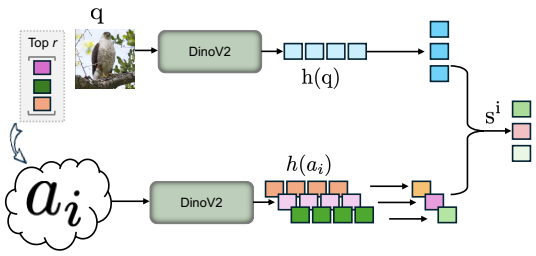


Figure 5: The visual representation of the re-ranker module.

K HUMAN EVALUATION

K.1 MATCHING DESCRIPTION WITH IMAGES

In this section, we analyze the discriminatory power of the species descriptions we generated through a qualitative human study. Annotators of the forced-choice matching test also classified each of the 200 tasks as easy, medium, or hard based on the perceived time and effort required.

As illustrated in Figure 6, easy tasks typically featured descriptions with a visual trait unique to the correct species, making the correct image immediately stand out. Tasks classified as medium, shown in Figure 7, required more focused attention because their key distinguishing features were more subtle. For hard tasks, shown in Figure 8, the distractors were extremely similar to the correct image, often differing by a single, minor attribute that required users to zoom in to discern the difference.

In Figure 9, we look at the cases where the annotators are inconclusive about their final decision, and in Figure 10, where annotators made incorrect choices. These scenarios occur under similar conditions, such as when a key visual trait is obscured in the image or an annotator overlooks a critical detail in the description. For instance, in the first row of Figure 9, the annotator states that a clearer image of one of the confusing options would greatly solve the issue. And, in the second row of Figure 10, the annotator incorrectly matches a bird because they mistook the ”pale crown stripe” as the white line above the bird’s eyes, forgetting that the pale strip is different as evident from the next sentence in description that tells us the pale stripe complements the white line and this is a feature present only in the correct image.

Ultimately, the difficulty of this fine-grained matching task shows the high quality and discriminatory power of our generated descriptions. They consistently include the specific, and often subtle, attributes necessary to differentiate species even in visually challenging scenarios.

The Bobolink (Dolichonyx oryzivorus) is a small New World blackbird distinguished by its unique plumage and size. Adult males are primarily black with striking creamy napes and white patches on their scapulars, lower backs, and rumps, making them easily recognizable. In contrast, adult females are mostly light brown with black streaks on their backs and flanks, and they have dark stripes on their heads, with darker wings and tails. Bobolinks measure between 5.9 to 8.3 inches (15–21 cm) in length, have a wingspan of about 10.6 inches (27 cm), and possess short, finch-like bills. These distinctive features, along with their unique coloration, help in identifying them from other similar bird species.



User: Primarily black removes last option, and only fourth one has a creamy nape, rest contradict. So I chose fourth.

The Geococcyx, commonly known as the roadrunner, is a distinctive bird found in the southwestern and south-central United States, Mexico, and Central America. It is a large, slender ground bird measuring 56 to 61 cm (22 to 24 in) in length, with a weight ranging from 230 to 430 g (8 to 15 oz). The roadrunner is characterized by its black-brown and white-streaked plumage, a prominent head crest, long legs, strong feet, and an oversized dark bill. Its tail is broad with white tips on the three outer feathers, and it has a unique bare patch of skin behind each eye, shaded blue in the front and red in the back. The roadrunner's zygodactyl feet leave distinct "X" track marks. Although capable of flight, it prefers to run, reaching speeds of up to 32 km/h (20 mph). In flight, its short, rounded wings reveal a white crescent in the primary feathers. These features, along with its unique vocalizations and habitat preferences, make the roadrunner easily identifiable.



User: Black brown and white-streaked excludes third, long legs excludes fourth and tail with white tips excludes fifth, the x shape of foot excludes first and second has clear head crest as well. So, I select second.

Figure 6: Examples of description matching where the annotators classified the cases as easy. Wrong options are highlighted in red, and the correct option is highlighted in green.

The Barn Swallow (Hirundo rustica) is a small passerine bird easily identifiable by its distinctive physical features. It measures 17–19 cm in length, including its elongated outer tail feathers, and has a wingspan of 32–34.5 cm. The adult male showcases steel blue upperparts with a rufous forehead, chin, and throat, separated from its off-white underparts by a broad dark blue breast band. Its most striking feature is the deeply forked "swallow tail" adorned with white spots across the outer end. Females resemble males but have shorter tail streamers, less glossy blue upperparts, and paler underparts. Juveniles are browner with a paler rufous face and lack the long tail streamers. The Barn Swallow's combination of a red face and blue breast band distinguishes it from similar species, such as the African Hirundo species and the Welcome Swallow.



User: Steel blue is only true for first and second, so rest are removed. First is clearly blue but the second has also bluish color on its back, it also has rufous chin throat which first does not, so I select second.

The Bohemian Waxwing (Bombycilla garrulus) is a starling-sized bird, measuring 19–23 cm in length with a wingspan of 32–35.5 cm. It is characterized by its buff-grey plumage, a prominent pointed crest, and distinctive black face markings. The wings are notable for their striking patterns, featuring white and bright yellow with some feathers tipped in red, resembling sealing wax. The tail is short, ending in a bright yellow band bordered by black. The bird's beak is mainly black, and its legs are dark grey or black. Both males and females look similar, though females have slightly less pronounced markings. Juveniles are duller, with fewer red wing tips and a smaller black face mask. The Bohemian Waxwing can be distinguished from the smaller Cedar Waxwing by its size and more vivid wing patterns, and from the Japanese Waxwing by the absence of a red tail band and different wing markings. Its call is a high trill, and in flight, it resembles a common starling with its fast, direct movement.



User: black face marking only present in first, second and third. The third option wing are only yellow and black and no white, so correct is either one or two. The second has no yellow on the wing, so first option should be correct.

Figure 7: Examples of description matching where the annotators classified the cases as medium. Wrong options are highlighted in red, and the correct option is highlighted in green.

The Common Tern (Sterna hirundo) is a medium-sized seabird distinguished by its slender build, long pointed wings, and deeply forked tail. Breeding adults exhibit pale grey upperparts and very pale grey underparts, with a striking black cap on the head. Their legs are orange-red, and the bill is narrow and pointed, varying in color from mostly red with a black tip to entirely black, depending on the subspecies. The upper wings are pale grey, developing a distinctive dark wedge as the summer progresses. In non-breeding adults, the forehead and underparts turn white, and the bill becomes all black or black with a red base. Juveniles have pale grey upper wings with a dark carpal bar and a ginger forehead that fades to white by autumn. The Common Tern can be differentiated from similar species like the Arctic Tern by its larger head, thicker neck, longer legs, and more triangular wings, as well as its more powerful flight. The Arctic Tern, in contrast, has greyer underparts and a consistently white wing throughout the summer. The Common Tern's tail does not extend beyond the wingtips when standing, unlike the Arctic and Roseate Terns. These features are crucial for identifying the Common Tern in photographs.



User: All are grey, but third has black legs so I remove it, second does not have black tip, forth has clear white underparts and not pale gray. It is either first or fifth, but given the color of fifth's beak and legs, first one is a much clearer match.

The Fish Crow (Corvus ossifragus) is a small crow species native to the eastern and southeastern United States, often found in wetland habitats. It is similar in appearance to the American Crow but is distinguishable by its smaller size, with a body length of 36–40 cm and a weight range of 247–320 grams. The Fish Crow has a silkier, smoother plumage with a blue or blue-green sheen on the upperparts and a greenish tint on the underparts. Its bill is somewhat slimmer, and it may have a small sharp hook at the end of the upper bill. When walking, Fish Crows appear to have shorter legs, and they tend to hunch and fluff their throat feathers when calling. The most distinctive feature is its call, a nasal "ark-ark-ark" or "nyuh unh," which contrasts with the American Crow's "caw caw." These vocal differences are often the most reliable way to differentiate between the two species.



User: First and fifth do not show any bluish or greenish color, the fourth only shows brownish color but not any blue or green, so its either second or third. The bill of third seems slimmer but it is not clear, the legs of second seem larger than third and also has bigger beak, I feel the the colors in third image are not clear but all other images have some contradictions, so I select the third one.

Figure 8: Examples of description matching where the annotators classified the cases as hard. Wrong options are highlighted in red, and the correct option is highlighted in green.

The Common Raven (Corvus corax) is a large, all-black bird distinguished by its impressive size, with a length ranging from 54 to 71 cm (21 to 28 inches) and a wingspan of 116 to 153 cm (46 to 60 inches). It is one of the heaviest passerines, weighing between 0.69 to 2.25 kg (1.52 to 4.96 lbs). The raven's plumage is mostly iridescent black, with a dark brown iris and elongated, pointed throat feathers. Its large, slightly curved beak is one of the largest among passerines, and it has a long, wedge-shaped tail. In flight, ravens are recognized by their stable soaring style and the creaking sound of their feathers. They are often confused with crows but can be distinguished by their larger size, heavier beak, shaggy throat feathers, and distinct wedge-shaped tail. The raven's deep, resonant croak is also unique, setting it apart from the calls of other corvids.



User: First one has white spot so its wrong, the third has no iridescent and no pointed throat feather, the second also lacks pointed throat feathers. The fifth is not clear to see if it has iridescent or not, so between fourth and fifth, they either have all features and no contradiction or low visibility. A closer image of the fifth one would help in selecting a single option.

The Cedar Waxwing (Bombycilla cedrorum) is a medium-sized bird, measuring about 6-7 inches in length with a wingspan of 8.7 to 11.8 inches. It is distinguished by its silky, shiny plumage in shades of brown, gray, and lemon-yellow, complemented by a subtle crest and a distinctive black mask bordered with white. A key identifying feature is the presence of small, red, wax-like droplets on the tips of its secondary flight feathers. The tail is typically yellow or orange, influenced by diet, and is short and square-tipped. The bird's beak is short and wide, and adults have a pale yellow belly. Immature birds are streaked and lack the adult's black mask. Cedar Waxwings are often seen in flocks, and their flight is strong and direct. They are known for their high-pitched whistles and buzzy trills.



User: Third and fourth have a black line and not a mask, so it is one of first second or fifth. The red drop is not visible in either, probably its hidden in the image. Both first and second have a yellow part on the tail but fifth does not. The mask is bordered by white on first, in the second image, I am not sure if that color is white or is it the beak but there is also some white near the eyes. I am leaning more towards first but not sure.

Figure 9: Examples of description matching where the annotators are unable to pick one single option. Wrong options are highlighted in red, and the correct option is highlighted in green.

Frigatebirds, belonging to the family Fregatidae, are large seabirds found across tropical and subtropical oceans. They are characterized by their predominantly black plumage, long, deeply forked tails, and long, slender hooked bills. A distinctive feature of male frigatebirds is their red gular pouch, which they inflate during the breeding season to attract females. Females, on the other hand, have white underbellies. Frigatebirds have long, narrow wings that taper to points, with a wingspan that can reach up to 2.3 meters, giving them the largest wing area to body weight ratio of any bird. Their wings and tails create a distinctive 'W' silhouette in flight. These birds are known for their aerial prowess, rarely flapping their wings and instead soaring on wind currents. They do not settle on the ocean surface due to their minimal oil production, which would cause their feathers to become waterlogged. Frigatebirds are also known for their kleptoparasitic behavior, occasionally stealing food from other seabirds.



User: Since all options dont have the red pouch, the correct option should be a female. All images except second show black plumage, but the second image since it is not whole, it still can not be discarded. Fifth and first do not show white underbellies, so they can be discarded, even the fourth option has yellowish-white. The third option is in flight but not creating a W but an M, so I think the second one to be the correct.

The Clay-colored Sparrow (Spizella pallida) is a small sparrow native to North America, easily identifiable by its distinct plumage and markings. It features light brown upperparts with darker streaks on the back and pale underparts. A key characteristic is its pale crown stripe set against a dark brown crown, complemented by a white line above the eyes and a dark line through the eyes. The bird also has a light brown cheek patch and brown wings with noticeable wing bars. Its short bill is pale with a dark tip, and the back of its neck is gray. The Clay-colored Sparrow has a long tail, which adds to its distinctive silhouette. These features help distinguish it from similar species like the Chipping Sparrow and Brewer's Sparrow, especially outside the breeding season when they may flock together.



User: All the images show some form of brown upperparts with some darker streaks. The crown is not clearly visible in first, third and fifth option, but the fourth image has a brownish crown and a white line above the eyes also with a dark line passing through the eyes. So, I think the fourth option is the correct one.

Figure 10: Examples of description matching where the annotators pick the wrong option.

K.2 Reliability Methodology

The primary goal of our inter-annotator reliability analysis was to quantify the consistency of human judgments on the quality of the generated descriptions. To provide a direct and interpretable measure of consensus among the annotators, we use Quadratic Weighted Agreement (QWA), which is a direct application of the weighting principle from Cohen (1968), simplified by not correcting for chance agreement. The approach is described in Algorithm 1.

Our dataset is characterized by a high prevalence of high-quality samples, as evident from the high average score, resulting in a skewed distribution of scores where lower ratings (e.g., 1 or 2 on a 5-point scale) are infrequent. In such cases, traditional chance-corrected reliability metrics like Cohen's Kappa or Krippendorff's Alpha can yield paradoxically low values Feinstein & Cicchetti (1990). This is not necessarily due to a lack of agreement, but because the high prevalence of a single rating makes it difficult to distinguish true agreement from chance agreement statistically.

To provide a direct and interpretable measure of consensus that is robust to this skew, we use Quadratic Weighted Agreement (QWA), which is a direct application of the weighting principle from Cohen (1968), simplified by not correcting for chance agreement. The approach is described in Algorithm 1.

Algorithm 1 Calculation of Quadratic Weighted Agreement (QWA)

Input: A set of items I, where each item $i \in I$ has a list of scores $R_i = [s_1, s_2, ..., s_n]$ from n annotators. The scoring scale has a minimum value S_{min} and a maximum value S_{max}.
Output: The overall Quadratic Weighted Agreement score, $QWA_{overall}$.

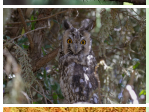
Initialize an empty list, A_{items}, to store the agreement score for each item.
$D_{max} \leftarrow (S_{max} - S_{min})^2$ ▷ Calculate max squared difference once
for all item i in I **do**
    Let R_i be the list of scores for item i.
    Initialize an empty list, W_{pairs}, to store weights for the current item.
    Let P_i be the set of all unique pairs of scores from R_i.
    for all pair (s_a, s_b) in P_i **do**
        $D_{pair} \leftarrow (s_a - s_b)^2$
        $W_{pair} \leftarrow 1 - \frac{D_{pair}}{D_{max}}$ ▷ Quadratic weighting formula
        Add W_{pair} to the list W_{pairs}.
    end for
    if W_{pairs} is not empty **then**
        $A_{item} \leftarrow$ average(W_{pairs}) ▷ Mean agreement for the item
        Add A_{item} to the list A_{items}.
    end if
end for
$QWA_{overall} \leftarrow$ average(A_{items}) ▷ Grand mean across all items
return $QWA_{overall}$

L Qualitative Description Samples

We present qualitative samples, including images of species and their descriptions, from our data. We show samples for various birds in Figure 11, and Figure 12. For our curated Pokémon dataset, we show examples in Figure 13.



Abbott's babbler (Malacocincla abbotti) is a small, nondescript bird measuring 12–13 cm in length, characterized by its short tail and heavy bill. Its plumage is primarily drab olive-brown, with a grayish-white throat and breast, and a white belly center. The flanks and undertail coverts are a distinctive rusty color, while the supercilium and lores are pale gray. Juveniles have darker rufescent-brown crowns and upperparts. Both sexes appear similar. This species is often found in low vegetation near streams, favoring areas with tree ferns and tangled undergrowth. Its distinctive calls, which include a series of rich, fluty notes, aid in identification. The subspecies M. a. krishnarajui is noted for its darker russet tail and rump compared to the nominate subspecies. Abbott's babbler is typically seen foraging in pairs close to the ground.



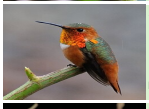
The Abyssinian owl, also known as the African long-eared owl (Asio abyssinicus), is a medium-sized owl distinguished by its dark brown eyes and black bill. It features gray eyebrows and prominent dark brown ear-tufts with white edges, which are slightly centrally located on its head. This owl is darker in appearance compared to the long-eared owl (Asio otus), with which it does not share a range. The Abyssinian owl's strong claws allow it to prey on a variety of animals, including smaller birds, field mice, and shrews. It inhabits open grasslands or moorlands with oak or cedar forests, typically found in mountain valleys and gorges at elevations up to 3,900 meters.



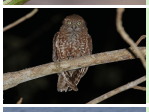
The Afep pigeon (Columba unicincta), also known as the African wood-pigeon or gray wood-pigeon, is a medium-sized bird measuring 35 to 36 cm in length and weighing between 356 and 490 grams. It is characterized by its gray neck and body, with darker gray wings and tail, and a distinctive white throat and belly. The breast displays a buff-pink hue. Notably, the Afep pigeon has red eyes and orbital rings, which are key features for identification. Its call is loud and distinctive, often described as "doo doo doo" or "whu whu whu whu-WHU." These pigeons are native to the African tropical rainforest regions flanking the Dahomey Gap.



The African blue flycatcher, also known as the blue-crested flycatcher (Elminia longicauda), is a small, dainty bird distinguished by its bright blue plumage and long tail. It features a short crest and its upper parts, including the tail, are a vibrant blue that transitions between blue and cyan. The bird's lores and flight feathers are black, with the latter edged in blue, while its underparts are a greyish blue that fades to whitish on the belly. The bill and legs are black, adding to its distinctive appearance. Juveniles are duller, with faint greyish spotting on the head and wing coverts. Measuring 15–18 cm in length and weighing 7–12 g, this flycatcher is often seen foraging gracefully with a fanned tail and half-open wings. These features, along with its unique coloration, make it easily identifiable among other bird species.



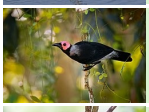
Allen's hummingbird (Selasphorus sasin) is a small, vibrant bird measuring 3 to 3.5 inches in length and weighing 2 to 4 grams. The male is distinguished by its green back and forehead, with rust-colored (rufous) flanks, rump, and tail, and an iridescent orange-red throat. In contrast, females and immature birds are mostly green with rufous only on the tail, which has white tips, and lack the iridescent throat patch, instead having speckled throats. A key distinguishing feature of the adult male Allen's hummingbird is the absence of a notch in the second rectrix, which helps differentiate it from the similar rufous hummingbird. Allen's hummingbirds are known for their energetic courtship displays and aggressive territorial behavior. They are primarily found along the coastal regions from California to Oregon, with some populations migrating to central Mexico for the winter.



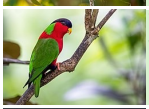
The Alor boobook (Ninox plesseni) is a small owl native to the Pantar and Alor Islands in the eastern Lesser Sunda Islands. It is distinguished by its compact size and rounded head, typical of the boobook owls. The plumage is predominantly brown with intricate patterns of lighter and darker streaks, providing excellent camouflage against tree bark. Its facial disc is less pronounced compared to other owls, with subtle white markings around the eyes. The eyes are large and dark, set against a relatively small, hooked beak. The wings are rounded, aiding in silent flight, and the tail is short. These features, along with its unique vocalizations, help differentiate the Alor boobook from other similar species.



The Chinese egret (Egretta eulophotes) is a medium-sized bird, averaging 68 cm in height, with distinctive all-white plumage throughout its life, similar to the little egret. Key identifying features include a dusky bill with a tannish peach base and yellow-green lores and legs outside the breeding season, while the iris remains yellow. During the breeding season, the Chinese egret develops a striking luxuriant crest over 11 cm long, along with long lanceolate breast plumes and dorsal aigrettes extending beyond the tail. The bill turns a bright orange-yellow, the lores become bright blue, and the legs turn black with yellow feet. These seasonal changes, particularly the vibrant breeding plumage and unique crest, help distinguish it from other egret species.



The Coleto (Sarcops calvus) is a distinctive bird species belonging to the family Sturnidae. It is characterized by its medium size and unique physical features that make it easily identifiable. One of the most striking characteristics of the Coleto is its bare, pinkish head, which contrasts sharply with its glossy black plumage. The bird's beak is short and stout, adapted for its varied diet. Its wings are rounded, aiding in agile flight through its forested habitat. The Coleto's overall appearance, with its combination of a bald head and sleek black body, sets it apart from other similar bird species, making it a unique subject for birdwatchers and photographers alike.



The collared lory (Vini solitaria) is a vibrant parrot species endemic to Fiji, easily identifiable by its striking plumage. Measuring approximately 20 cm (7.9 in) in length, this bird features bright red undersparts and face, a dark purple crown, and greenish upperparts. The nape is lime green with some elongated red feathers, and the lower abdomen is purple. Its bill is yellow-orange, feet are pink-orange, and irises are orange-red. Males and females are similar, though females have a paler crown with a greenish hue towards the back. Juveniles are duller with vague purple striations on the upper abdomen and breast, and they have a brown beak and pale brown irises. The collared lory's adaptation to urban environments, such as Suva, and its fast, straight flight with quick shallow wingbeats further distinguish it.



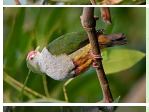
The collared sunbird (Hedydipna collaris) is a tiny bird, measuring only 9\u201310 cm (3.5\u20133.9 in) in length, and is part of the Nectariniidae family. It is characterized by its short, thin, down-curved bill and brush-tipped tubular tongue, adaptations for nectar feeding. The adult male is distinguished by its glossy green upperparts and head, a yellow belly, and a narrow purple breast band. In contrast, the female is a duller green above and entirely yellow below. These birds have a fast and direct flight on short wings and are typically found in forests near water across sub-Saharan Africa. The collared sunbird's unique combination of size, bill shape, and vibrant plumage makes it identifiable from other similar species.



The Comoro black parrot (Coracopsis sibilans) is a medium-sized parrot distinguished by its predominantly dark plumage, which appears blackish with subtle brownish tones. This bird features a robust, slightly curved beak that is well-suited for its diet. Its wings are relatively broad and rounded, aiding in its flight through the dense forests of the Comoros, where it is endemic. The parrot's tail is moderately long and squared off at the end. Unlike many other parrots, the Comoro black parrot lacks the bright, vivid colors often associated with the family, making its dark, muted coloration a key identifying feature.



The collared finchbill (Spizixos semitorques) is a distinctive songbird known for its unique physical characteristics. It features a striking black head and a contrasting white collar around its neck, which is a key identifier. The bird's body is predominantly olive-green, providing a vibrant contrast to its darker head. It has a stout, finch-like beak that is well-suited for its frugivorous diet. The wings are rounded, and the tail is relatively short, aiding in its agile flight through forested hills. This species is typically found in China, Taiwan, Japan, and Vietnam, and is often seen in moderate elevation forests. The collared finchbill's combination of a black head, white collar, and olive-green body makes it easily distinguishable from other similar species.



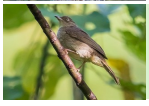
The Beautiful Fruit Dove (Ptilinopus pulchellus), also known as the rose-fronted pigeon or crimson-capped fruit dove, is a small bird measuring approximately 19 cm (7.5 inches) in length. It is primarily green with distinctive features that aid in its identification: a striking red crown, a whitish throat, a greenish-yellow bill, and purplish-red feet. Its breast is blue-grey, transitioning to a yellowish-orange belly, with a notable reddish-purple patch in between. Both male and female birds share these vibrant color patterns. This species is native to the rainforests of New Guinea and nearby islands in West Papua, Indonesia. The Beautiful Fruit Dove's unique combination of colors and small size make it distinguishable from other fruit doves.



The common scimitarbill (Rhinopomastus cyanomelas) is a distinctive bird characterized by its dark blue plumage, which can appear almost black in certain lighting. A key feature for identification is its long, curved beak, which is black in adults and grey in juveniles. This bird's sleek, elongated body and scimitar-shaped bill set it apart from other species. It is typically found across various regions in Africa, including Angola, Botswana, and South Africa, among others. When identifying this bird from a photograph, look for its unique combination of dark blue coloration and the pronounced curve of its beak.



The Corsican finch (Carduelis corsicana) is a small bird in the true finch family, Fringillidae, distinguished by its unique plumage and coloration. It features dark-streaked brown upperparts, which contrast with its brighter yellow underparts, a key characteristic that sets it apart from the similar citril finch. The Corsican finch has a typical finch-like beak, suited for seed eating, and its overall size is small, consistent with other finches. This bird is endemic to the Mediterranean islands, including Corsica, Sardinia, Elba, Capraia, and Gorgona, which can also aid in its identification.



The cream-eyed bulbul (Pycnonotus pseudosimplex) is a distinctive bird endemic to Borneo, notable for its unique cream-colored eyes, which set it apart from other bulbuls. This species is generally found in old-growth hill forests, emphasizing its habitat specificity. It shares a close genetic relationship with the ashy-fronted bulbul of Palawan, rather than the red-eyed Bornean population of the cream-vented bulbul. The cream-eyed bulbul's plumage and overall appearance are similar to other bulbuls, but the eye color is a key distinguishing feature for identification.

Figure 11: Examples of bird species with their descriptions.



Figure 12: Examples of more bird species with their descriptions.

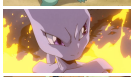
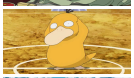
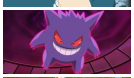
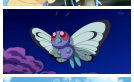
	Pikachu is a small, chubby, and agile rodent-like Pokèmon with bright yellow fur. It has large, pointed ears with black tips and a small, triangular nose. Its eyes are round and black, with white pupils, and it has a distinctive red circle on each cheek. Pikachu's tail is shaped like a lightning bolt, with a brown base. It stands on two short legs, and its forearms are small, with five fingers. Pikachu's overall appearance is both cute and energetic, embodying its Electric-type nature.
	Bulbasaur is a small, quadrupedal Pokèmon with a teal, amphibian-like body. Its skin is dotted with darker green patches. A large, green plant bulb grows on its back, which is the source of its Grass-type abilities. Bulbasaur has red eyes, a wide mouth, and pointed ears. Its legs are short and sturdy, each ending in three sharp claws. The bulb on its back is crucial for photosynthesis, contributing to its growth and evolution. Bulbasaur's appearance is both endearing and distinctive, making it easily recognizable among Pokèmon species.
	Charizard is a large, dragon-like Pokèmon with a fiery orange body and blue-green wings. Standing on two powerful legs, it has a long, tapering tail ending in a flame that reflects its mood and health. Its head features a blunt snout, sharp fangs, and piercing blue eyes, with two horn-like structures on top. Charizard's broad wings and muscular arms suggest strength, while its belly and the underside of its tail are cream-colored. Its overall imposing stature and fiery tail make it easily recognizable as a formidable Fire Flying-type Pokèmon.
	Squirtle is a small, turtle-like Pokèmon with a smooth, light blue body and a round shape. It has large, expressive eyes and a slightly pointed tail. Its most distinctive feature is its brown shell, which is segmented with a pale yellow underside. Squirtle's limbs are short and stubby, with three-clawed digits on each hand and foot. Its head is large in proportion to its body, and its mouth often forms a playful grin. This Water-type Pokèmon is known for its ability to withdraw into its shell for protection.
	Mewtwo is a bipedal, humanoid Pokèmon with a sleek, muscular build. Its skin is predominantly light gray with a long, purple tail and a similarly colored tube extending from the back of its skull to its spine. It has a feline-like face with piercing, angular eyes and small, pointed ears. Mewtwo's hands are three-fingered, and its feet are large with two prominent toes. Standing at roughly 6'7", it exudes a powerful, imposing presence, embodying both strength and intelligence.
	Eevee is a small, quadruped Pokèmon with a primarily brown fur coat and a bushy, cream-colored tail tip. It has large, expressive brown eyes and long, pointed ears with dark inner coloring. Its face is adorned with a fluffy, cream-colored ruff around its neck, adding to its endearing appearance. Eevee's compact body and short legs give it a nimble, agile stance. Its unique genetic makeup allows for numerous evolutionary paths, reflected in its adaptable look. Eevee's charming and versatile appearance makes it easily recognizable and beloved among Pokèmon enthusiasts.
	Jigglypuff is a small, round Pokèmon with a soft, pink body resembling a balloon. Its most distinctive features are its large, expressive blue eyes and a tuft of curly hair on its forehead. Jigglypuff has short, stubby arms and feet, and its ears are pointed with dark interiors. It often carries a microphone, reflecting its ability to sing lullabies that induce sleep. This Pokèmon's cute, spherical shape and cheerful demeanor make it easily recognizable.
	Ditto is a small, amorphous Pokèmon with a gelatinous, pink body that resembles a blob of putty. It has a simplistic face marked by two small, beady eyes and a wide, straight-lined mouth. Its body is highly flexible, allowing it to transform into the shape of other Pokèmon. Typically about a foot tall, Ditto lacks distinct limbs, giving it a smooth, featureless appearance. Its ability to mimic any Pokèmon's form makes it unique, though its transformations often retain Ditto's signature facial expression, providing a subtle clue to its true identity.
	Psyduck is a bipedal, yellow Pokèmon resembling a duck with a plump body and webbed feet. It features a rounded head with a pair of vacant, staring eyes and a small, flat bill. Notably, three tufts of black hair sprout from its head. Its arms are short and often seen clutching its head, a gesture linked to its frequent headaches. Psyduck's tail is short and stubby, complementing its overall pudgy appearance. Standing at about 2 feet 7 inches, its unassuming demeanor belies its latent psychic abilities.
	Snorlax is a large, rotund Pokèmon, predominantly teal-blue with a cream-colored face, belly, and feet. Its body resembles a giant bear with a round shape and short, stubby limbs. Snorlax has small, pointed ears and closed, slanted eyes, often giving it a sleepy appearance. Its arms are short but thick, ending in sharp, white claws. Its massive size and relaxed posture, usually seen lying down or sitting, make it easily recognizable. The contrast between its dark body and light underbelly, along with its serene expression, are key identifiers.
	Gengar is a Ghost/Poison-type Pokèmon with a round, shadowy body and a mischievous grin. Its purple skin is smooth, and it has pointed ears and short, spiky protrusions on its back. Gengar's eyes are red and sinister, adding to its eerie appearance. It stands at a height of about 4 feet 11 inches, with stubby arms and legs, each ending in three clawed digits. Gengar's presence often casts a shadowy aura, and it seems to hover slightly above the ground, enhancing its spectral, ghost-like qualities.
	Vaporeon is an aquatic, quadrupedal Pokèmon with a sleek, blue body reminiscent of a mermaid. It features a distinctive, finned tail similar to a fish, and its neck is adorned with a white, frilled collar. Vaporeon has a small, cat-like face with large, expressive dark eyes and pointed ears. Its head is crowned with a fin-like crest, and it has webbed feet, enhancing its swimming abilities. This elegant Pokèmon is known for its ability to blend seamlessly with water, thanks to its translucent, water-like texture.
	Arcanine is a majestic, canine Pokèmon with a striking orange coat adorned with bold black stripes. It has a large, flowing mane and bushy tail, both cream-colored, giving it a regal appearance. Its face features pointed ears, a black nose, and sharp, confident eyes. Standing at an impressive height, Arcanine exudes power and grace. Its muscular build and long limbs suggest speed and agility, while its proud posture reflects its legendary status. The tufts of fur around its legs and neck enhance its noble and fiery aura, making it easily recognizable.
	Gyarados is a massive, serpentine Pokèmon with a fierce demeanor. Its body is primarily blue with a white underbelly, segmented by ridged scales. It boasts intimidating red eyes and a wide, gaping mouth filled with sharp fangs. A prominent, jagged crest adorns its head, flanked by long, flowing barbels. Its dorsal and tail fins are large and spiky, adding to its fearsome appearance. Gyarados typically measures around 21 feet in length, exuding an aura of raw power and aggression, often depicted with its mouth open in a roar.
	Butterfree is a butterfly-like Pokèmon with a predominantly white body and large, vivid purple wings adorned with black vein-like patterns. Its eyes are large and red, providing excellent vision. It has two slender, black antennae and a small, blue face. Its limbs are thin, with three-fingered hands and two-toed feet. Butterfree's wings are dusted with scales that can release a toxic powder. Standing at about 1.1 meters tall, its delicate yet vibrant appearance and gentle fluttering motion make it easily recognizable.
	Lapras is a large, sea-dwelling Pokèmon resembling a plesiosaur. It features a gentle expression, a long neck, and a rounded, blue body marked with darker blue spots. Its most distinctive feature is the gray, shell-like carapace on its back, adorned with blunt knobs. Lapras has a smooth, white underside, and its flippers are paddle-like, aiding in swimming. Its head is adorned with a pair of small, spiral-shaped ears and intelligent, kind eyes. Standing at over eight feet tall, Lapras exudes a calm, majestic presence, often seen ferrying passengers across bodies of water.
	Blastoise is a large, bipedal turtle-like Pokèmon with a robust blue body and a cream-colored underside. It stands out with its powerful, tank-like appearance, accentuated by two water cannons protruding from its shell. The shell is brown with a white rim, providing both defense and a base for its hydro cannons. Blastoise has a large, rounded head with a pair of small, pointed ears and fierce red eyes. Its limbs are strong and stubby, ending in three clawed digits. The overall impression is one of strength and resilience, embodying its role as a formidable Water-type Pokèmon.",
	Scizor is a sleek, bipedal Pokèmon with a metallic red exoskeleton. It stands out with large, pincer-like claws that resemble scythes, each marked with a black dot that mimics an eye. Its body is streamlined, featuring a narrow waist and segmented abdomen. Scizor's wings are thin and transparent, hinting at its Bug-type heritage. Its head is angular with sharp, pointed eyes and small, pointed horns. Scizor's legs are sturdy, ending in pointed feet, and its overall appearance exudes strength and agility. This Steel/Bug-type Pokèmon is known for its striking color and formidable presence.
	Luxray is a sleek, feline Pokèmon with predominantly black fur and a striking blue mane. It stands on four legs, featuring sharp, yellow eyes and a distinctive star-shaped tail tip. Its ears are pointed with blue interiors, and its face is marked by a blue, mask-like pattern. Luxray's body is muscular, with a pronounced chest and tufts of fur on its elbows. Its mane and tail are edged with electric blue, highlighting its Electric-type nature. This Pokèmon exudes a powerful and commanding presence, making it easily recognizable.
	Ninetales is a graceful, fox-like Pokèmon with a sleek, pale golden coat and nine luxurious, flowing tails tipped with a lighter hue. Its slender body and pointed snout give it an elegant appearance. Ninetales has sharp, intelligent red eyes and large, pointed ears. The mane around its neck is slightly ruffled, adding to its regal demeanor. Standing at about 3'7" tall, its presence is both majestic and mysterious. Each of its tails is said to possess mystical powers, often shimmering with a subtle glow, reflecting its mystical and enchanting nature.
	Metagross is a large, metallic Pokèmon resembling a robotic arachnid. It has a blue, disc-shaped body with four sturdy legs, each ending in sharp, silver claws. Its face features a prominent, metallic "X" across the front, giving it a formidable appearance. Metagross's eyes are red and piercing, adding to its intimidating look. The Pokèmon's overall color is a steely blue, with silver accents on its claws and face. Its massive size and distinctive shape make it easily recognizable, exuding an aura of strength and intelligence.

Figure 13: Pokémon species with their descriptions.