# Divide and Conquer: Structured Reranking for Expert-Level Ecological Image Retrieval

Asmi Kumar Edward Vendrow Sara Beery
Massachusetts Institute of Technology
{asmi,evendrow,beery}@mit.edu

## Abstract

*Fine-grained image retrieval in scientific domains such as ecology demands compositional and expert-level reasoning that general-purpose VLMs often lack. In this paper, we introduce a two-stage structured reranking pipeline that augments queries with web-sourced expert knowledge and decomposes them into verifiable subquestions using large multimodal models. Then, images are scored against these subqueries to produce a final relevance ranking. Our method consistently boosts retrieval accuracy, especially on behavioral and contextual queries, while also greatly reducing manual effort, improving interpretability, and advancing automated visual search for scientific research.*

## 1. Introduction

Image retrieval plays a central role in many real-world tasks, from everyday search to scientific analysis. In recent years, large-scale vision-language models (VLMs) such as CLIP [24] have made significant progress in zero-shot image retrieval. Trained on hundreds of millions of image-text pairs from publicly available online sources, these models can retrieve relevant images for a diverse range of natural language queries [28]. They perform especially well when queries describe common concepts or scenes that are frequently represented or similarly expressed in training data [24, 28].

However, zero-shot VLMs often struggle with queries that require fine-grained semantic understanding beyond broad scene recognition. They frequently fail on tasks involving compositional or relational reasoning, such as semantic role reversals ("a horse eating grass" vs. "grass eating a horse"), negation ("cat on mat" vs. "cat not on mat"), and spatial relations ("dog behind tree" vs. "tree behind dog") [1, 18, 25, 32]. Benchmarks like ARO highlight these weaknesses by generating negative captions that swap attributes or word order, showing that state-of-the-art models misinterpret such distinctions [45].

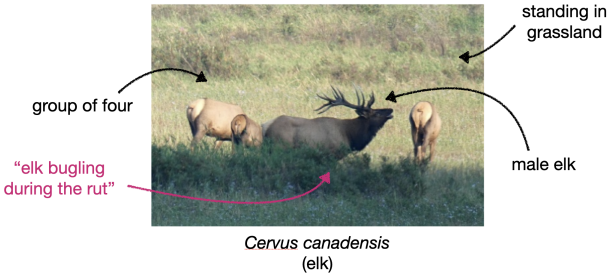



Figure 1. An expert wants to search iNaturalist for *"elk bugling during the rut"*. When images are labeled with only a species name, retrieving such specific behavioral or contextual information becomes challenging, and we miss valuable secondary data.

These compositional and relational reasoning failures are especially problematic in scientific image search, where queries combine multiple specialized concepts and require detailed visual evidence. In ecology, for example, researchers rely on large image datasets to document behaviors, monitor individuals, and assemble training data [26, 42]. Unlike generic image search for "a bird," scientists must answer research-driven questions. An ornithologist might query *"juvenile and adult birds in courtship display"* [35], requiring identification of both age classes and specific behaviors. A conservationist searching for *"a California condor tagged with green 26"* [35] must locate a rare species and read small numbered tags on moving animals, demanding precise reasoning across multiple visual and semantic cues.

Meeting these information needs depends on the availability of large, well-annotated image datasets. Platforms like iNaturalist host over 250 million user-contributed biodiversity images [11], but most are minimally labeled, typically only with a species name, timestamp, and GPS location [34]. As a result, rich "secondary data" [22]—such as behavior, habitat, or physical context—remains unannotated and difficult to access, as shown in Figure 1. Extracting this information at scale is labor-intensive and requires expert review [42]. Unfortunately, general-purpose models

often fail to interpret expert-level queries, returning results that may appear superficially relevant but miss the scientific intent [23, 32, 41, 45]. A system capable of understanding and responding to complex queries would unlock significant value for ecological research and dataset curation.

### 1.1. Reranking

Addressing expert information needs necessitates advanced retrieval strategies. Real-world pipelines in both document [15, 19] and image retrieval [36, 44] often employ a two-stage process: an inexpensive initial retrieval of a broad candidate set, followed by a more sophisticated but costlier *reranking* of the top results, a subset of the candidates. The reranking stage is where there is still significant room for improvement, especially in specialized domains [27, 43]. While initial retrieval aims to identify relevant items from the entire dataset, reranking refines a smaller set of candidates. This refinement is critical in expert contexts, where relevance often depends on fine-grained, domain-specific cues. Reranking enables deeper reasoning over a limited set of plausible items, incorporating context and structured analysis that would be impractical to apply at full scale [48]. For expert users, this mirrors real-world search workflows: scientists often begin with an approximate result set and iteratively refine it using more detailed criteria.

Despite the narrowed candidate space of the reranking stage, performance remains limited in expert-level tasks. For instance, the recent INQUIRE benchmark, discussed in Section 4, has shown that even state-of-the-art multimodal large language models (LLMs) achieve mean average precision below 50% on their reranking tasks [35], falling substantially short of a perfect 100% ranking.

### 1.2. Contributions

We address two main challenges in expert-level image retrieval: the need for specialized background knowledge and the requirement for fine-grained, compositional reasoning over visual content [38, 40]. Our key insight is that expert queries are often implicitly composed of multiple visual subquestions, each of which could be evaluated independently [3, 14, 47]. We propose a novel two-stage pipeline for expert-level image reranking that:

1. *Enriches the LLM context with background knowledge* by retrieving relevant textual information from a web search that provides scene description and clarifies specialized or ambiguous terminology.
2. *Decomposes complex queries* by prompting an LLM to generate discrete yes/no subqueries, each targeting a distinct aspect of the original complex query (e.g., species identity, behavior, or environmental context).
3. *Performs evaluation* using the vision capabilities of OpenAI's GPT-4 proprietary models to evaluate each candidate image against the generated subqueries.

Subquery-level predictions are aggregated into a final relevance score, which we use to rerank the initial retrievals. We find that our approach yields consistent improvements in retrieval accuracy. In addition to performance gains, this structured reranking procedure provides interpretable intermediate outputs, making the model's decision process more transparent to the user.

## 2. Related Work

**VLMs for Retrieval.** Modern text-to-image retrieval systems build upon three decades of research in multimodal representation learning and information retrieval. The INQUIRE benchmark [35] represents a recent advancement in expert-level retrieval evaluation for natural world queries. Early contrastive learning approaches like CLIP [24] established the paradigm of joint vision-language embedding spaces, enabling zero-shot retrieval through cosine similarity comparisons. Subsequent work improved alignment quality through scaled training [5] and noise-resistant objectives [46], with models like ALIGN [13] demonstrating the value of web-scale pretraining. Domain-specific adaptations have proven particularly valuable in ecological applications, with BioCLIP [30] and WildCLIP [9] extending foundation models to taxonomic recognition tasks. The iNaturalist dataset [34] has served as a critical testbed for these developments, with its 2024 iteration providing unprecedented scale for training and evaluation [35].

**Reranking in Text-to-Image Retrieval.** Reranking has emerged as a critical component in retrieval pipelines, especially for expert-level queries where initial retrievals may be noisy or insufficiently precise. This process has been increasingly adopted in the vision-language domain, where transformer-based rerankers refine initial candidate sets by modeling global and local feature interactions [31]. The INQUIRE benchmark introduced a dedicated reranking task (INQUIRE-RERANK) to isolate and evaluate this stage, showing that advanced models like GPT-4V and GPT-4o can substantially improve over CLIP's initial rankings, yet still are far from manual expert-level performance [35].

Recent work has explored the use of LLMs as rerankers, leveraging their reasoning capabilities to assess candidate relevance more holistically. Iterative query refinement through conversations with LLMs improves reranking [16], while knowledge-enhanced reranking with MLLMs [6] incorporates external information, paralleling our expert context injection (Section 3.1). Nonetheless, reranking remains a bottleneck for expert-level ecological queries.

**Structured Reranking for Expert Queries.** Structured reasoning pipelines have emerged as a promising strategy for tackling the compositional complexity of expert-level

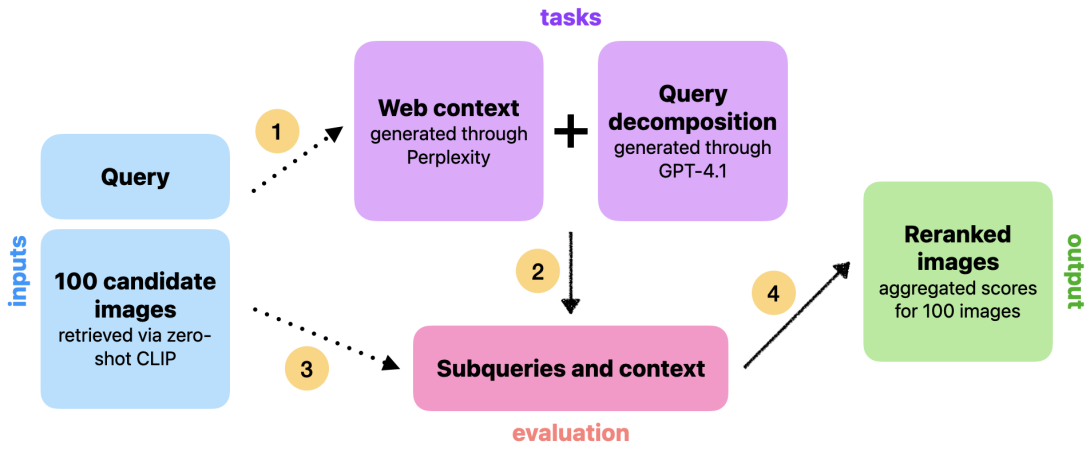



Figure 2. Overview of our reranking pipeline. The query is decomposed with retrieved context. Then, each candidate image is then assessed on the subqueries. These per-subquery decisions are aggregated into a final score that produces the reranked list of images.

image retrieval. In ecological domains, where queries often involve fine-grained species identification, behavioral context, and visual attributes, traditional retrieval models frequently fall short. The INQUIRE benchmark [35] illustrates these limitations: while reranking with large VLMs like GPT-4o significantly improved performance over CLIP-based baselines, the best model (GPT-4o) still achieved only 47.1 mean average precision. Performance degraded further on semantically dense queries, particularly those requiring morphological precision or domain-specific terminology.

However, recent advancements in multimodal reasoning highlight the benefits of decomposing complex queries into simpler, visually verifiable components [14]. In retrieval-augmented generation (RAG), OmniSearch [2] represents a self-adaptive retrieval agent that dynamically plans each retrieval action based on the question's solution stage and current retrieval content. This approach reflects human-like reasoning by decomposing complex questions into sub-question chains [40]. Relatedly, ViLBench [33] uses supervised reward models to provide detailed step-wise feedback.

The Multi-Attribute Composition (MAC) benchmark [39] further underscores the challenges faced by models when handling objects with multiple interrelated attributes. With its comprehensive attribute annotations, MAC emphasizes the need for models that can reason about nuanced, multi-attribute compositions.

Applying structured reasoning strategies to ecological domains is particularly promising. Ecological queries often involve intricate combinations of species identification, behavioral context, and environmental attributes. By integrating generated expert context, decomposing queries into explicit subcomponents, and using structured prompt sequences, our approach aligns more closely with the hierarchical validation processes used by domain experts.

## 3. Methodology

We propose a structured reranking pipeline that augments the original query with expert textual context, decomposes it into a small set of visually verifiable subqueries, and uses a multimodal model to assess a candidate image against these subqueries. The resulting subquery-level predictions are aggregated into a final relevance score. This score is then used to re-order the initial set of $k$ retrieved candidate images (which contains a mix of relevant and irrelevant results), with the goal of placing the relevant images at the top of the re-ranked list. Figure 2 overviews this process.

### 3.1. Expert Context Injection

In fine-grained image retrieval tasks, a primary challenge lies in bridging the knowledge gap between general-purpose vision models and expert-level understanding of specialized concepts. To bridge this gap, we introduce *expert context injection*, a retrieval-augmented approach that enriches queries with relevant domain-specific context. This injects essential background knowledge directly into the LLM's context for accurate subquery generation and answering.

For each query, we use Perplexity's web search API, Sonar, to fetch a concise explanatory passage that defines key terms and highlights distinguishing visual characteristics. For example, for the query *"Water frogs in amplexus position"*, the retrieved Perplexity context might clarify that amplexus refers to "a mating position in frogs and toads where the male grasps the female from behind, fertilizing eggs as they are released."

This context is prepended to all prompts in our pipeline when querying GPT-4.1. We use GPT-4.1 because it offers a larger context window, improved instruction-following capabilities, and lower costs compared to models like GPT-

**Algorithm 1** Expert-Guided Reranking Pipeline

**Require:** Query $q$; candidate image set $\mathcal{I} = \{I_1, \ldots, I_K\}$

$C_q \leftarrow$ PERPLEXITYWEBSEARCH$(q)$ ▷ Retrieve expert context.
$S_q \leftarrow$ LLM_GENERATESUBQUERIES$(q, C_q)$ ▷ Generate subqueries $\{s_1, \ldots, s_n\}$.
**for** each image $I_i \in \mathcal{I}$ **do**
  *history* $\leftarrow []$ ▷ Initialize Q&A history.
  **for** each subquery $s_j \in S_q$ **do**
    $(a_{ij}, p_{ij}) \leftarrow$ LLM_ANSWERQUERY$(s_j, I_i, q, C_q, \textit{history})$
    Append $(s_j, a_{ij})$ to *history*
  **end for**
  $S_i \leftarrow \frac{1}{n} \sum_{j=1}^{n} p_{ij}$ ▷ Aggregate confidence scores.
**end for**
**return** Images sorted in descending order of $S_i$

4o, making it especially effective for our workflow [21]. This technique draws on broader work in RAG, which has been shown to improve LLM performance on knowledge-intensive tasks without additional fine-tuning [12, 17]. Our use of GPT-4.1 is consistent with the INQUIRE baseline, where proprietary models substantially outperform open-source ones on ranking tasks [35].

### 3.2. Subquery Generation

Along with expert context, we decompose complex domain-specific queries into smaller components. This decomposition addresses a key challenge in expert-level image retrieval: queries often contain implicit compositional requirements that must all be satisfied simultaneously.

GPT-4.1 transforms an original query into 2-3 binary subqueries that collectively capture its semantic intent and follow a logical, hierarchical progression. For example, for the query *"Water frogs in amplexus position"*, the pipeline could first generate "Does this image show Water Frogs?" to confirm the subject's identity. Then, a second subquery would verify the specific behavior: "Are the water frogs in a position where one is on top of the other, gripping with its forelimbs?", leveraging expert context and converting abstract details into a visually verifiable format. Each subquery is a yes/no question answerable through visual evidence alone.

We limit subqueries to two or three based on empirical testing showing diminishing returns with additional questions. For further guidance, Appendix 8.1 contains a visualization of the subquery generation thought process.

### 3.3. Subquery Answering

Given a candidate set of images $\{I_1, I_2, \ldots, I_K\}$ retrieved by a base model (following the INQUIRE benchmark, we use CLIP ViT-H/14 [35]), we evaluate each image against the set of subqueries generated in the previous stage. Each subquery $s_j$ is a yes/no question designed to isolate a specific visual criterion derived from the original query $q$. Our goal is to determine, for each image $I_i$, whether it satisfies a specific subquery and with what level of confidence.

To do this, we use GPT-4.1's multimodal capabilities to assess answering each subquery $s_j$ with the image $I_i$, where $i = 1, \cdots, 100$. Each prompt includes the original query $q$, the expert context paragraph $C_q$, the subquery $s_j$, and the candidate image $I_i$. The model is instructed to answer with either "Yes" or "No," accompanied by a natural language reasoning explanation (i.e., using chain-of-thought). For scoring, however, we extract only the binary answer and the model's confidence in it.

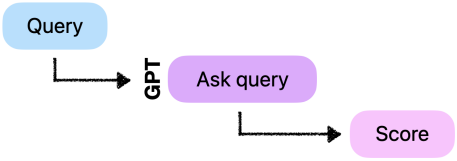
To quantify confidence, we draw inspiration from selective prediction work on calibrated VLM confidence estimates, which prompt models to self-verify their answers [29]. We adapt a similar idea by directly leveraging GPT-4.1's token-level output probabilities. Specifically, we extract the logits associated with the tokens "Yes" and "No" from the model's top-20 predictions and pass them through a softmax to yield a scalar confidence score.

$$p_{ij} = \frac{\exp(\text{logit}_{ij}^{(\text{Yes})})}{\exp(\text{logit}_{ij}^{(\text{Yes})}) + \exp(\text{logit}_{ij}^{(\text{No})})} \times 100,$$

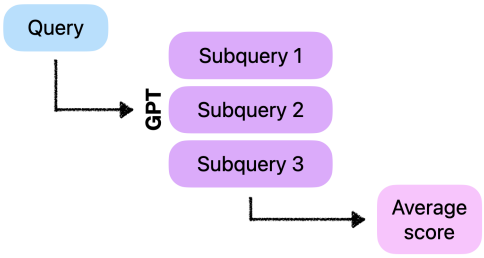
where $p_{ij} \in [0, 100]$ represents the model's confidence that image $I_i$ satisfies subquery $s_j$. A score close to 100 indicates high confidence in "Yes," while a score near 0 reflects confidence in "No." In cases where neither token appears in the top-20 logits of an individual subquery, we default to scoring as if the answer is "No" to minimize false positives.

### 3.4. Subquery Chaining and Final Reranking

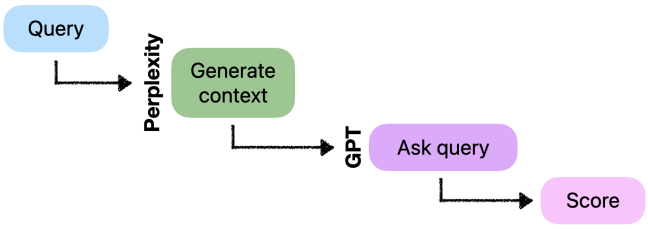
To leverage the inherent dependencies among subqueries, we implement a *chaining strategy*. When evaluating a given subquery $s_{j+1}$ on a specific image $I_i$, the LLM's context is augmented with the set of all previous subqueries (for the original query at hand) and their corresponding generated answers, $\{(s_k, a_{ik})\}_{k=1}^{j}$. Here, $s_k$ denotes the $k^{\text{th}}$ subquery and $a_{ik}$ represents the answer to subquery $s_k$ specifically



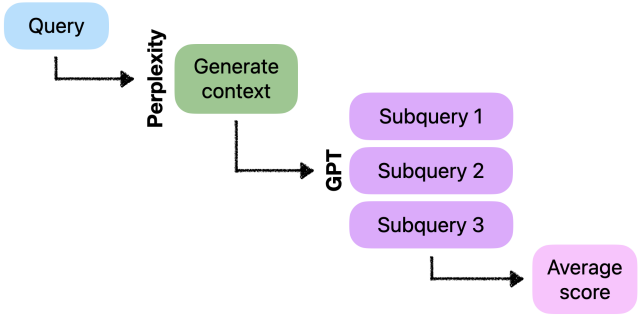

(a) **BASELINE (Direct Query)**: The model is asked, "Does this image show {query}? Answer with yes or no." No subqueries or additional background context are provided. This matches the original INQUIRE reranking method, after replacing the earlier GPT-4V with newer GPT-4 variants.




(b) **SUBQUERIES ONLY**: The model is prompted with a sequence of automatically GPT-4.1-generated subquestions (subqueries) that break the query into verifiable visual components. Each one is answered independently. No expert context is supplied.




(c) **CONTEXT ONLY**: The model receives an explanation describing the visual criteria necessary to satisfy the query, generated with Perplexity web search, but is asked only the original undecomposed query.




(d) **SUBQUERIES + CONTEXT (Full Model)**: The model is prompted with both the expert context and the full set of subqueries. This setting represents our complete reranking pipeline and allows the model to leverage structured guidance alongside domain-specific background knowledge.

Figure 3. Baseline and ablation setups.

for image $I_i$. This sequential context allows the model to build on prior information, reducing irrelevant reasoning by guiding its assessment (e.g., confirming presence of a bird before assessing if the bird is flying).

Each subquery response yields a confidence score $p_{ij}$. We then compute an overall relevance score for image $I_i$ as the mean of its $n$ subquery confidences:

$$S_i = \frac{1}{n} \sum_{j=1}^{n} p_{ij}.$$

Images that satisfy all subqueries achieve scores near 100, while any low-confidence answer lowers the average proportionally. Finally, we sort the candidates $\{I_i\}$ in descending order of $S_i$, producing a reranked list that places the most query-relevant images at the top. See Appendix 8.2 for a visual representation of top-$k$ ranking for a given query. The full inference-time pipeline is summarized in Algorithm 1.

# 4. Experimental Setup

## 4.1. Dataset and Queries

We evaluate our method on the INQUIRE benchmark, which provides a testbed for assessing reranking performance on challenging ecological queries. This benchmark was introduced to systematically evaluate the limitations of VLMs in expert settings. The INQUIRE-RERANK task, which we focus on, comprises 160 expert-authored queries. For each query, the benchmark includes a fixed set of 100 candidate images, which were initially retrieved using zero-shot CLIP ViT-H/14 [8] from the entire five-million-image iNaturalist 2024 dataset (iNat24) [35]. Domain experts have already classified each of these 100 candidate images as either relevant or not relevant to the query, focusing on ecologically meaningful content such as animal interactions, rare morphologies, and complex visual scenes.

The 160 queries span four semantic supercategories:

APPEARANCE, BEHAVIOR, CONTEXT, and SPECIES. We exclude the six SPECIES queries from our primary analysis due to insufficient sample size, focusing on the remaining 154 queries. To further characterize query difficulty, we introduce another axis of stratification based on linguistic complexity. Queries containing specialized scientific terminology (e.g., *frogs in amplexus*, *osmeterium eversion*) are manually designated as LINGO, while those phrased in more colloquial or accessible English (e.g., *caterpillar showing orange horns*) are designated as NON-LINGO. Table 1 summarizes the distribution.

| Supercategory | Total count | Non-lingo | Lingo |
|---|---|---|---|
| Appearance | 52 | 43 | 9 |
| Behavior | 59 | 54 | 5 |
| Context | 43 | 39 | 4 |
| **Total** | **154** | **136** | **18** |

Table 1. Number of queries in each supercategory, split by language complexity.

### 4.2. Reranking Experiments

For reranking, each query is paired with the same 100 CLIP-retrieved candidate images. We instantiate our framework using Perplexity to generate expert-informed context paragraphs and GPT-4.1 to perform subquery-based image evaluation. We evaluate multiple experimental variants of our method, as summarized in Figure 3.

## 5. Results and Discussion

Our evaluation of reranking methods on our 154 queries reveals category-dependent trends in retrieval performance, as well as effects of query decomposition and contextual augmentation. All results are reported as mean AP@100 with standard error, and are stratified by query supercategory (APPEARANCE, BEHAVIOR, CONTEXT) and linguistic complexity (LINGO vs. NONLINGO).

Figure 4 summarizes the overall performance of each reranking strategy. Our SUBQUERIES + CONTEXT method is the highest-performing, highlighting the synergy of our proposed techniques from Section 3.

Since the release of INQUIRE in late 2024, no retrieval methods beyond the original dataset paper have yet reported results on this benchmark. Our work marks the first exploration of structured reranking on INQUIRE and establishes an initial reference point for future research.

### 5.1. Stratification Rationale

We stratify queries by both supercategory and linguistic complexity to target areas where VLMs tend to underperform. Among the supercategories, we particularly look

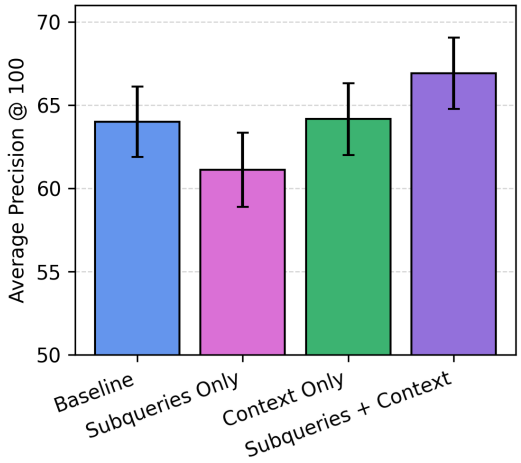



Figure 4. Mean AP scores across all queries for each reranking method. The SUBQUERIES + CONTEXT pipeline achieves the highest mean AP (66.9), outperforming the baseline (64.0), using subqueries alone (61.1), and using context alone (64.2).

to BEHAVIOR and CONTEXT for improvement, as these queries encode compositional scenes whose elements (e.g., specific behaviors, object interactions) are more likely to be represented and learned from the vast web-scale datasets on which pretrained multimodal models are trained. For example, we can decompose queries like *"A male and female cardinal sharing food"* into precisely defined components that models can more easily identify.

In contrast, APPEARANCE queries often hinge on subtle morphological differences such as color variants, markings, or sexual dimorphism (e.g., *"Tropical Orb Weaver with multiple white spots in the abdominal dorsum"*, *"adult male Misumena vatia"*). These distinctions demand fine-grained anatomical localization and taxonomic knowledge, presenting greater challenges for VLMs trained on general-purpose data.

The LINGO versus NON-LINGO categorization allows for further probing. This stratification enables us to assess whether reranking can provide improvements across both expert-facing (LINGO) and more common-language (NON-LINGO) queries.

### 5.2. Supercategories

To parse our results, we look at performance by query supercategory and linguistic complexity in Figures 5 and 6, respectively.

**CONTEXT Queries.** The joint SUBQUERIES + CONTEXT method yields the largest gains for CONTEXT queries, with an AP@100 of 76.5, compared to 71.1 for the baseline. Improvements are present for both NON-LINGO queries (+5.6 points) and LINGO queries (+3.8 points).

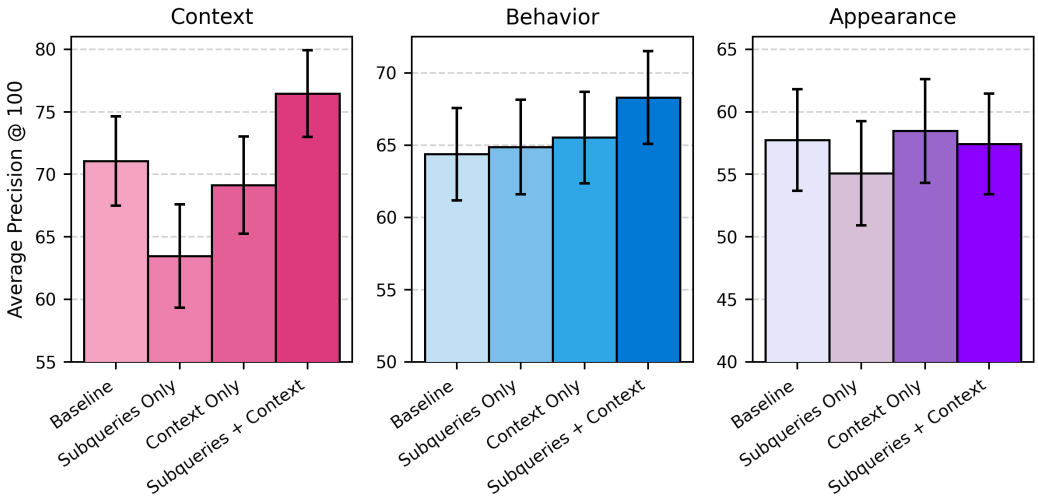



Figure 5. AP scores for each supercategory (CONTEXT, BEHAVIOR, APPEARANCE) across different experimental methods. Combining subqueries with expert-provided context consistently outperforms baselines for BEHAVIOR and CONTEXT, but not for APPEARANCE.

**BEHAVIOR Queries.** For BEHAVIOR queries, SUBQUERIES + CONTEXT also leads to a clear improvement (68.3), outperforming the Baseline (64.4). Notably, the benefit for LINGO (+5.1 points) exceeds that for NON-LINGO (+3.8 points) queries, reflecting the method's ability to resolve compositional and technical requirements by decomposing queries and grounding them in expert context. Error bars indicate some variability, but the gains are nonetheless significant.

**APPEARANCE Queries.** APPEARANCE queries are the most challenging, with all methods generally achieving lower scores and larger error bars than the other supercategories, particularly for LINGO queries. The blended SUBQUERIES + CONTEXT method did not outperform alternatives; CONTEXT ONLY improved performance by +6.3 points relative to the baseline. This suggests that subquery decomposition may have diluted query precision or introduced ambiguity, while context augmentation alone better preserved relevant cues.

### 5.3. Linguistic Complexity

When the full reranking pipeline is applied, LINGO queries in the CONTEXT and BEHAVIOR categories see the largest absolute gains. This indicates that structured decomposition and expert context are especially beneficial for semantically dense or ambiguous queries, where the model's default knowledge is insufficient. The technical or specialized language of LINGO queries often encodes multiple compositional requirements (e.g., species, behavior, context) or uses domain-specific terminology that general-purpose models struggle to interpret. Injecting expert context and decomposing queries into visually verifiable subcomponents allows the pipeline to resolve these ambiguities. While benefits for NON-LINGO queries are clear, they are not always as dominant.

Interestingly, NON-LINGO queries, typically phrased in accessible language, tend to underperform relative to LINGO ones, at least more so than one might expect (Figure 6). While NON-LINGO queries can appear simpler, their reduced technical precision often introduces ambiguity to the model. Consider the following real examples from our dataset:

- *"measuring the body dimensions of a bee"*: "Measuring" is an abstract human action, difficult for VLMs to visually verify without explicit context (e.g., a ruler, human hands using a ruler).
- *"A Eurasian Red Squirrel gathering material for its nest"*: "Gathering material" is a broad action lacking specific visual cues (e.g., the type of material, how it's carried, whether or not the nest must be visible in the image). Such general phrasing can lead to discrepancies between an expert's precise interpretation and a VLM's broader understanding.
- *"Mexican grass-carrying wasp visiting a flower"*: "Visiting" is a vague term, allowing for many visual interpretations that may not match the user's specific intent (e.g., landing, flying near or in front of, pollinating).

From the prior examples, it is evident that NON-LINGO queries can omit important details or use colloquial phrasing that fails to specify all relevant constraints, leading to a broader or less well-defined candidate set. As a result, even with structured reranking, the model may struggle to con-

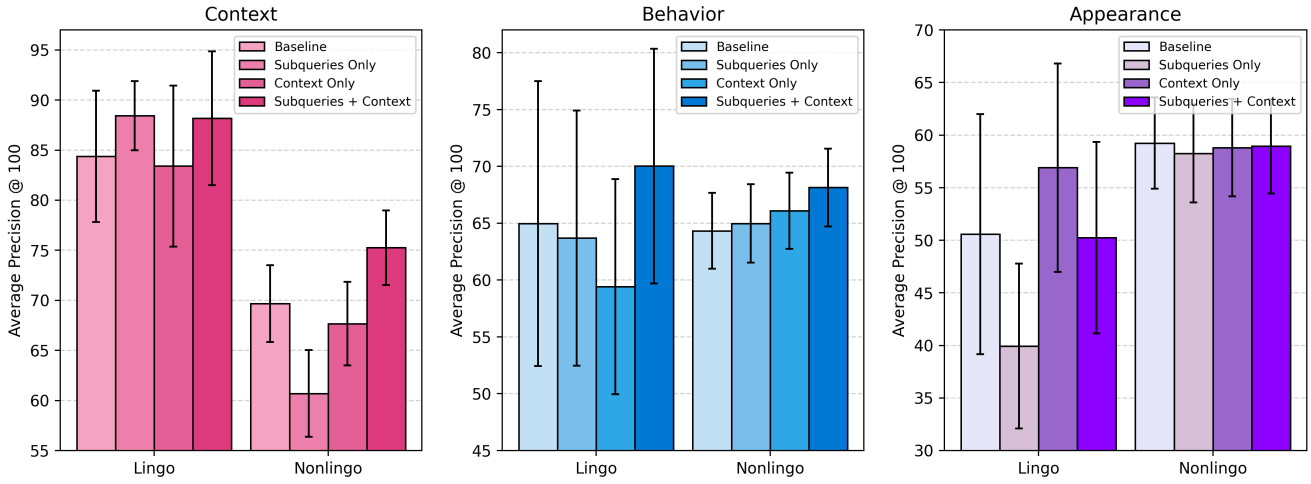



Figure 6. Mean AP scores for linguistic complexity for (left) CONTEXT, (middle) BEHAVIOR, and (right) APPEARANCE queries. Reranking improvement holds across query types, though the magnitude of gain depends on query features.

sistently identify the most relevant images, especially when multiple plausible interpretations exist.

### 5.4. Ablation Insights

**SUBQUERIES ONLY.** While introducing structure, this method can generate queries that miss crucial context or even contradict the query's intent. For example, for the query *"albino American robin"*, generated subqueries sometimes asked about an "orange-red breast"—a trait by definition absent in albino animals. Decomposing queries without expert context can also introduce noise, especially when the original query is already atomic or when visual discrimination is the primary challenge.

**CONTEXT ONLY.** Contextual augmentation alone generally matches or exceeds the baseline but rarely meets the performance of the SUBQUERIES + CONTEXT approach. Without explicit decomposition, the model must still reason over the entire query at once, making it more likely to miss compositional constraints or subtle distinctions that subqueries capture.

**SUBQUERIES + CONTEXT.** We directly address the weaknesses of the prior ablations by ensuring each subquery is grounded in expert knowledge through the generated context paragraph and targets the most relevant visual features, so that their collective satisfaction accurately reflects the image's overall relevance to the main query. Returning to *"albino American robin"*, with context added, subqueries correctly focus on white feathers and the presence of pink or red eyes.

## 6. Limitations and Impact

Our dataset size limits statistical power for rare query types. Subquery decomposition also assumes attributes of interest are visible; in practice, posture or occlusion could hide key traits. Consistent with literature, we find retrieval performance can be sensitive to query phrasing [37]. For further stability, a "3×" voting scheme (three evaluations per image with a majority vote) could be used.

Our study relies on proprietary models, as budgeting constraints limited broader model evaluation. Preliminary experiments showed smaller or alternative models (e.g., GPT-4.1-mini, -4o, -o1) underperformed or were less cost-effective, motivating our focus on GPT-4.1. Further comparisons remain important in future work.

By showing that structured subqueries with web-searched context deliver substantial improvements, this work builds a foundation for reliable automated visual retrieval in ecology and related fields. Still, over-reliance on imperfect context or prompts must be avoided; human oversight remains essential for ecological decision-making.

## 7. Conclusion

We introduced a two-stage reranking pipeline that decomposes wildlife queries into structured subquestions with scene-level context. This approach improves retrieval on compositional behavioral and contextual queries, though challenges remain for fine-grained appearance tasks. By improving reranking accuracy, our method reduces manual effort in behavioral annotation, habitat assessment, and dataset curation. Future work could explore richer context, adaptive prompting, and human-in-the-loop refinement to better support ecological research.

## References


[1] Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip HS Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29612–29622, 2025. 1

[2] Alibaba-NLP. Omnisearch: A self-adaptive retrieval agent for multimodal rag. https://github.com/Alibaba-NLP/OmniSearch, 2024. 3

[3] Tianyi Bai, Zengjie Hu, Fupeng Sun, Jiantao Qiu, Yizhen Jiang, Guangxin He, Bohan Zeng, Conghui He, Binhang Yuan, and Wentao Zhang. Multi-step visual reasoning with visual tokens scaling and verification. *arXiv preprint arXiv:2506.07235*, 2025. 2

[4] Oscar Beijbom, Tali Treibitz, David I Kline, Gal Eyal, Adi Khen, Benjamin Neal, Yossi Loya, B Greg Mitchell, and David Kriegman. Improving automated annotation of benthic survey images using wide-band fluorescence. *Scientific reports*, 6(1):23166, 2016. 2

[5] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 2

[6] Zhanpeng Chen, Chengjin Xu, Yiyan Qi, and Jian Guo. Mllm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. *arXiv preprint arXiv:2407.21439*, 2024. 2

[7] Ellen M Ditria, Christina A Buelow, Manuel Gonzalez-Rivero, and Rod M Connolly. Artificial intelligence and automated monitoring for assisting conservation of marine ecosystems: A perspective. *Frontiers in Marine Science*, 9: 918104, 2022. 2

[8] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023. 5

[9] Valentin Gabeff, Marc Rußwurm, Devis Tuia, and Alexander Mathis. Wildclip: Scene and animal attribute retrieval from camera trap data with domain-adapted vision-language models. *International Journal of Computer Vision*, 132(9): 3770–3786, 2024. 2

[10] Manuel Gonzalez-Rivero, Oscar Beijbom, Alberto Rodriguez-Ramirez, Dominic EP Bryant, Anjani Ganase, Yeray Gonzalez-Marrero, Ana Herrera-Reveles, Emma V Kennedy, Catherine JS Kim, Sebastian Lopez-Marcano, et al. Monitoring of coral reefs using artificial intelligence: A feasible and cost-effective approach. *Remote Sensing*, 12 (3):489, 2020. 2

[11] iNaturalist Team. We've reached 250 million verifiable observations! *iNaturalist Blog*, 2025. 1

[12] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 1(2):4, 2022. 4

[13] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[14] Chenchen Jing et al. Maintaining reasoning consistency in compositional visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 123–132, 2022. 2, 3

[15] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48, 2020. 2

[16] Saehyung Lee, Sangwon Yu, Junsung Park, Jihun Yi, and Sungroh Yoon. Interactive text-to-image retrieval with large language models: A plug-and-play approach. *arXiv preprint arXiv:2406.03411*, 2024. 2

[17] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020. 4

[18] Daniela Massiceti, Camilla Longden, Agnieszka Slowik, Samuel Wills, Martin Grayson, and Cecily Morrison. Explaining clip's performance disparities on data from blind/low vision users. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12172–12182, 2024. 1

[19] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. Multi-stage document ranking with bert. *arXiv preprint arXiv:1910.14424*, 2019. 2

[20] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25):E5716–E5725, 2018. 1, 2

[21] OpenAI. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/, 2025. 4

[22] Nadja Pernat, Susan Canavan, Marina Golivets, Jasmijn Hillaert, Yuval Itescu, Ivan Jarić, Hjalte MR Mann, Pavel Pipek, Cristina Preda, David M Richardson, et al. Overcoming biodiversity blindness: Secondary data in primary citizen science observations. *Ecological Solutions and Evidence*, 5 (1):e12295, 2024. 1

[23] Uwe Peters and Benjamin Chin-Yee. Generalization bias in large language model summarization of scientific research. *Royal Society Open Science*, 12(4):241776, 2025. 2

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 2

[25] Tim Rädsch, Leon Mayer, Simon Pavicic, A Emre Kavur, Marcel Knopp, Barış Öztürk, Klaus Maier-Hein, Paul F Jaeger, Fabian Isensee, Annika Reinke, et al. Bridging vision language model (vlm) evaluation gaps with a framework for scalable and cost-effective benchmark generation. *arXiv preprint arXiv:2502.15563*, 2025. 1

[26] Robert Sanders. Using ai and inaturalist, scientists build one of the highest resolution maps yet of california plants. 2024. 1

[27] Yash Saxena, Anpur Padia, Mandar S Chaudhary, Kalpa Gunaratna, Srinivasan Parthasarathy, and Manas Gaur. Ranking free rag: Replacing re-ranking with selection in rag for sensitive domains. *arXiv preprint arXiv:2505.16014*, 2025. 2

[28] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 1

[29] Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Raghavi Chandu. Selective" selective prediction": Reducing unnecessary abstention in vision-language reasoning. *arXiv preprint arXiv:2402.15610*, 2024. 4

[30] 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, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19412–19424, 2024. 2

[31] Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. Instance-level image retrieval using reranking transformers. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 12105–12115, 2021. 2

[32] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 1, 2

[33] Haoqin Tu, Weitao Feng, Hardy Chen, Hui Liu, Xianfeng Tang, and Cihang Xie. Vilbench: A suite for vision-language process reward modeling. *arXiv preprint arXiv:2503.20271*, 2025. 3

[34] 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*, pages 8769–8778, 2018. 1, 2

[35] Edward Vendrow, Omiros Pantazis, Alexander Shepard, Gabriel Brostow, Kate Jones, Oisin Mac Aodha, Sara Beery, and Grant Van Horn. Inquire: A natural world text-to-image retrieval benchmark. *Advances in Neural Information Processing Systems*, 37:126500–126514, 2024. 1, 2, 3, 4, 5

[36] Yabing Wang, Zhuotao Tian, Qingpei Guo, Zheng Qin, Sanping Zhou, Ming Yang, and Le Wang. From mapping to composing: A two-stage framework for zero-shot composed image retrieval. *arXiv preprint arXiv:2504.17990*, 2025. 2

[37] Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, et al. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*, 2024. 8

[38] Ran Xu, Wenqi Shi, Yuchen Zhuang, Yue Yu, Joyce C Ho, Haoyu Wang, and Carl Yang. Collab-rag: Boosting retrieval-augmented generation for complex question answering via white-box and black-box llm collaboration. *arXiv preprint arXiv:2504.04915*, 2025. 2

[39] Shuo Xu, Sai Wang, Xinyue Hu, Yutian Lin, Sibei Yang, and Yu Wu. Mac: A benchmark for multiple attributes compositional zero-shot learning. *arXiv preprint arXiv:2406.12757*, 2024. 3

[40] Shangzi Xue, Zhenya Huang, Jiayu Liu, Xin Lin, Yuting Ning, Binbin Jin, Xin Li, and Qi Liu. Decompose, analyze and rethink: Solving intricate problems with human-like reasoning cycle. *Advances in Neural Information Processing Systems*, 37:357–385, 2024. 2, 3

[41] Yibo Yan, Shen Wang, Jiahao Huo, Jingheng Ye, Zhendong Chu, Xuming Hu, Philip S Yu, Carla Gomes, Bart Selman, and Qingsong Wen. Position: Multimodal large language models can significantly advance scientific reasoning. *arXiv preprint arXiv:2502.02871*, 2025. 2

[42] Chih-Hsuan Yang, Benjamin Feuer, Zaki Jubery, Zi K Deng, Andre Nakkab, Md Zahid Hasan, Shivani Chiranjeevi, Kelly Marshall, Nirmal Baishnab, Asheesh K Singh, et al. Arboretum: A large multimodal dataset enabling ai for biodiversity. *arXiv e-prints*, pages arXiv–2406, 2024. 1

[43] Tetiana Yemelianenko, Iuliia Tkachenko, Tess Masclef, Mihaela Scuturici, and Serge Miguet. Learning to rank approach for refining image retrieval in visual arts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1623–1631, 2023. 2

[44] Shuhei Yokoo, Kohei Ozaki, Edgar Simo-Serra, and Satoshi Iizuka. Two-stage discriminative re-ranking for large-scale landmark retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1012–1013, 2020. 2

[45] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022. 1, 2

[46] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 2

[47] Qinggang Zhang, Hao Chen, Junnan Dong, Wentao Li, Feiran Huang, and Xiao Huang. Structure-guided large language models for text-to-sql generation. 2025. 2

[48] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*, 2023. 2

# Divide and Conquer: Structured Reranking for Expert-Level Ecological Image Retrieval

Supplementary Material

## 8. Appendix

### 8.1. Subquery Strategy

Our approach to subquery generation begins by identifying the species or object of interest. We then formulate *visually verifiable* questions that can be discerned from images alone, ensuring that each subquery confirms an image's relevance to the main query. Figure 7 walks through a sample thought process.

### 8.2. Reranking Visualization

For reranking, we begin with 100 candidate images. Taking the query *"a nest with eggs displaying brood parasitism by a cowbird"* as an example, we generate a context paragraph and subqueries. Each candidate is evaluated for binary relevance, and images are then ranked. See Figure 8 for a visualization.

### 8.3. Runtime and Cost Analysis

Table 2 summarizes the estimated token counts and costs for a single query in our pipeline. Each query involves 300 GPT-4.1 answering calls (100 candidates $\times$ 3 subqueries, an upper bound), one subquery-generation call, and one Perplexity Sonar call to retrieve web-searched context. Costs represent an upper bound per query. With reasonable batching across candidate images, the full pipeline runs in under 15 seconds per query.

| Stage | Calls | Input | Output | Cost ($) |
|---|---|---|---|---|
| Perplexity Sonar | 1 | ∼20 | ∼100 | 0.0001 |
| GPT-4.1 Subq. Gen. | 1 | ∼1,000 | ∼60 | 0.0025 |
| GPT-4.1 Subq. Ans. | 300 | ∼121,500 | ∼600 | 0.2478 |
| **Total** | 302 | ∼140,500 | ∼850 | **0.2504** |

Table 2. Estimated token usage and cost per query. Pricing: GPT-4.1 at $2/M input and $8/M output; Perplexity Sonar at $1/M input and output.

While the per-query cost is higher than that of traditional search methods, the trade-off lies in the significant time saved. Labeling a six-month batch of Snapshot Serengeti images with species, counts, and basic behaviors—so they could later be searched by scientists—has required 2 to 3 months of work by thousands of volunteers [20]. Even after datasets are labeled (if they are labeled with such detail at all), ecologists often spend hours sifting through search results, discarding irrelevant images, and verifying matches. At scale, this burden quickly becomes overwhelming. In contrast, our system reduces this effort to seconds, pro-

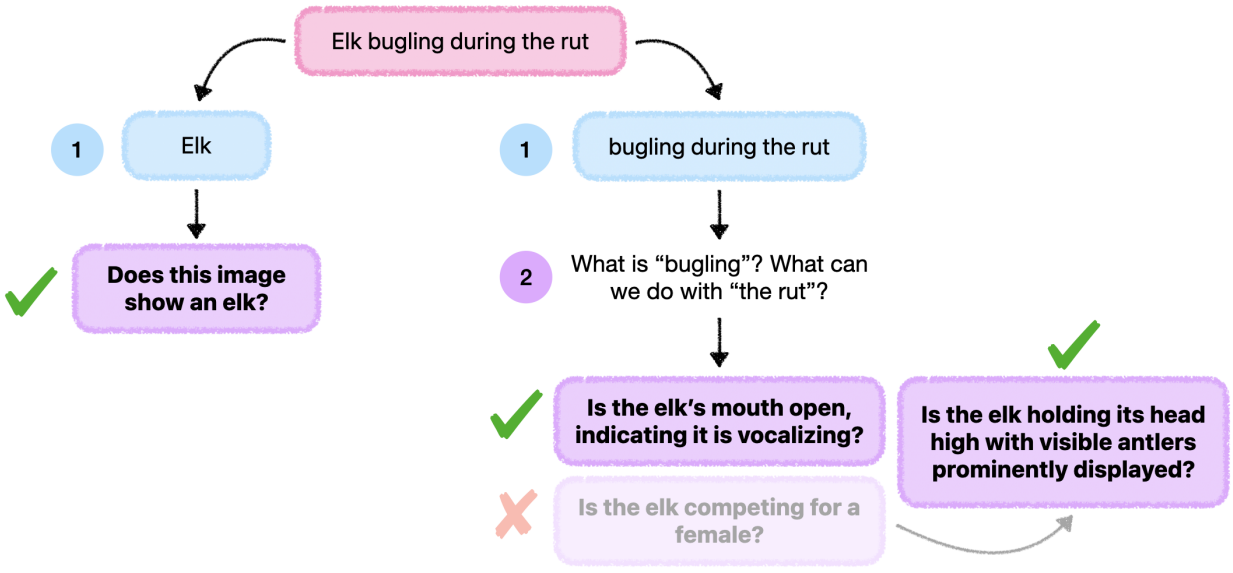
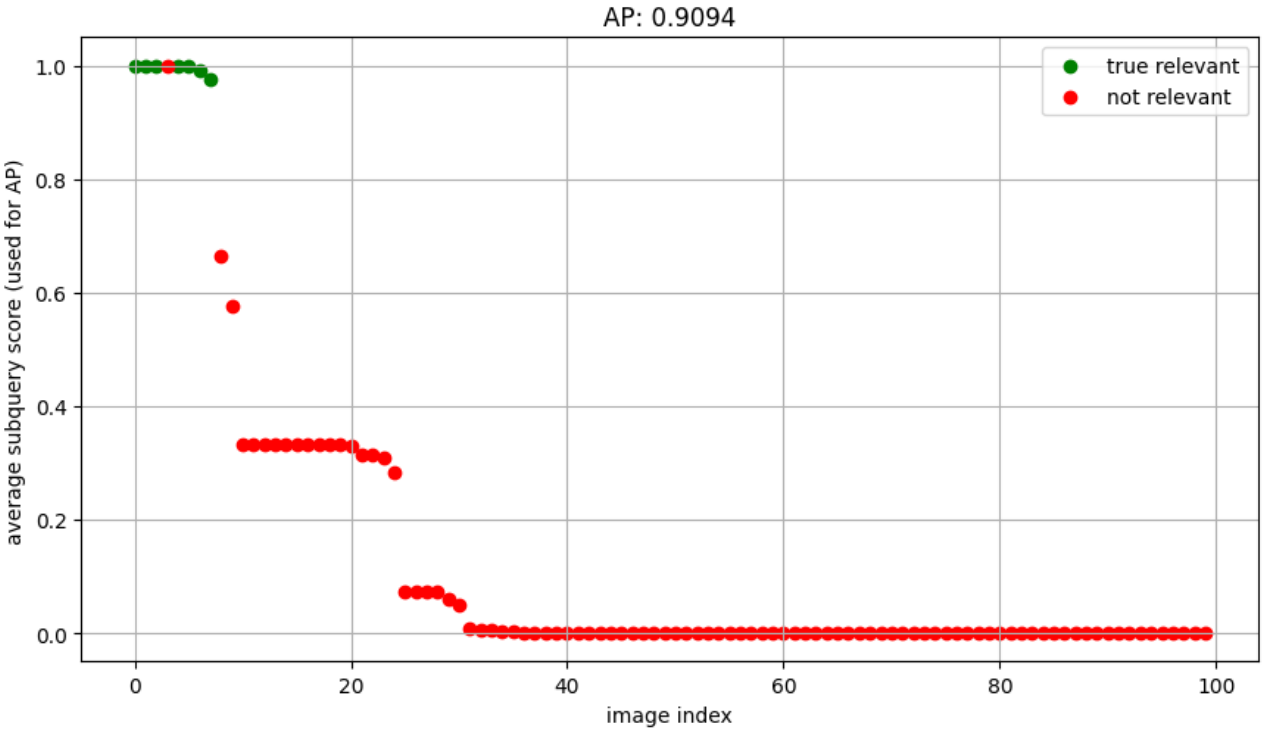



Figure 7. General approach to structuring the subquery generation prompt. The process begins by identifying the species or object of interest, followed by formulating visually verifiable questions that can be answered from images alone.




| | |
|---|---|
| Query | a nest with eggs displaying brood parasitism by a cowbird |
| Subquery 1 | Does this image show a cowbird or its distinctive eggs in the nest? Answer with yes or no. |
| Subquery 2 | Are there eggs in the nest that are visibly different in appearance from the others (e.g., in size, color, or markings)? Answer with yes or no. |
| Subquery 3 | Is there a chick or egg that appears significantly larger or otherwise distinct compared to the rest? Answer with yes or no. |

Figure 8. Visualization of reranking for the query *"a nest with eggs displaying brood parasitism by a cowbird"*. Relevant images (green) are clustered near the top, while non-relevant images (red) follow. The stepwise ranking pattern reflects how many subqueries an image satisfies: those at the top confidently satisfy all three, while those at the bottom satisfy none. Subqueries are rarely answered ambiguously, which is why relatively few images appear between the "steps."

ducing a ranked set of high-confidence results that can be used immediately. This enables researchers to move quickly from curation to scientific inquiry, such as analyzing behavior, testing hypotheses, or planning fieldwork, without bottlenecked annotation or searching. In practice, the modest computational cost is far outweighed by the speed and efficiency gains, and the time saved can ultimately lower overall project costs [4, 7, 10, 20].

### 8.4. Prompts and Code

Code and LLM propmts for this paper are available at:
github.com/asmikumar/rerank-ecological-images