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-

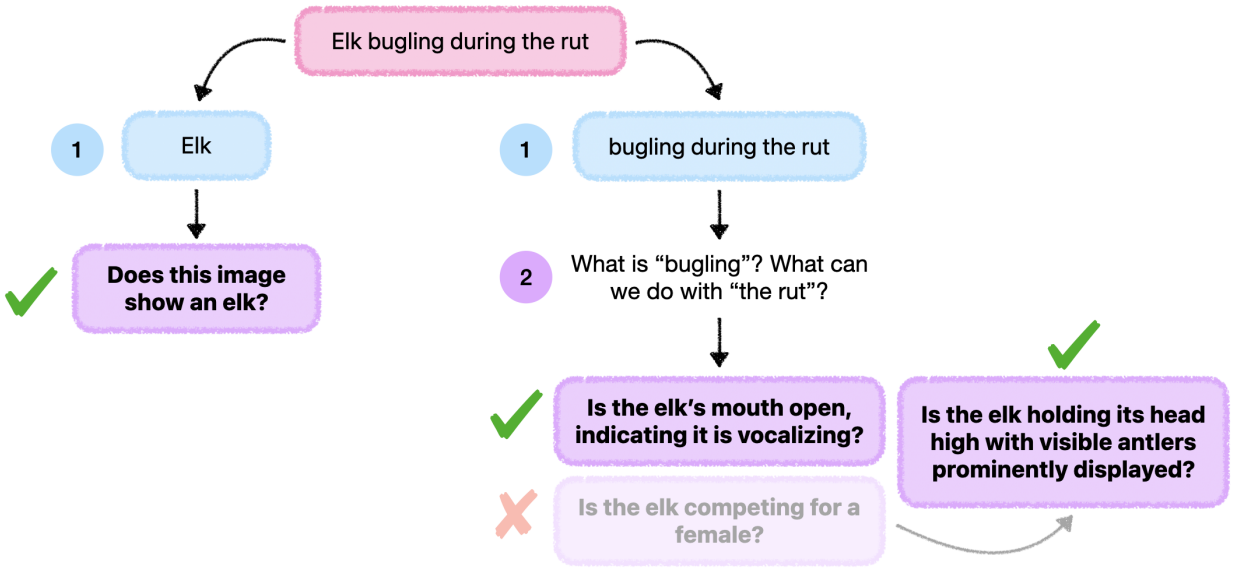
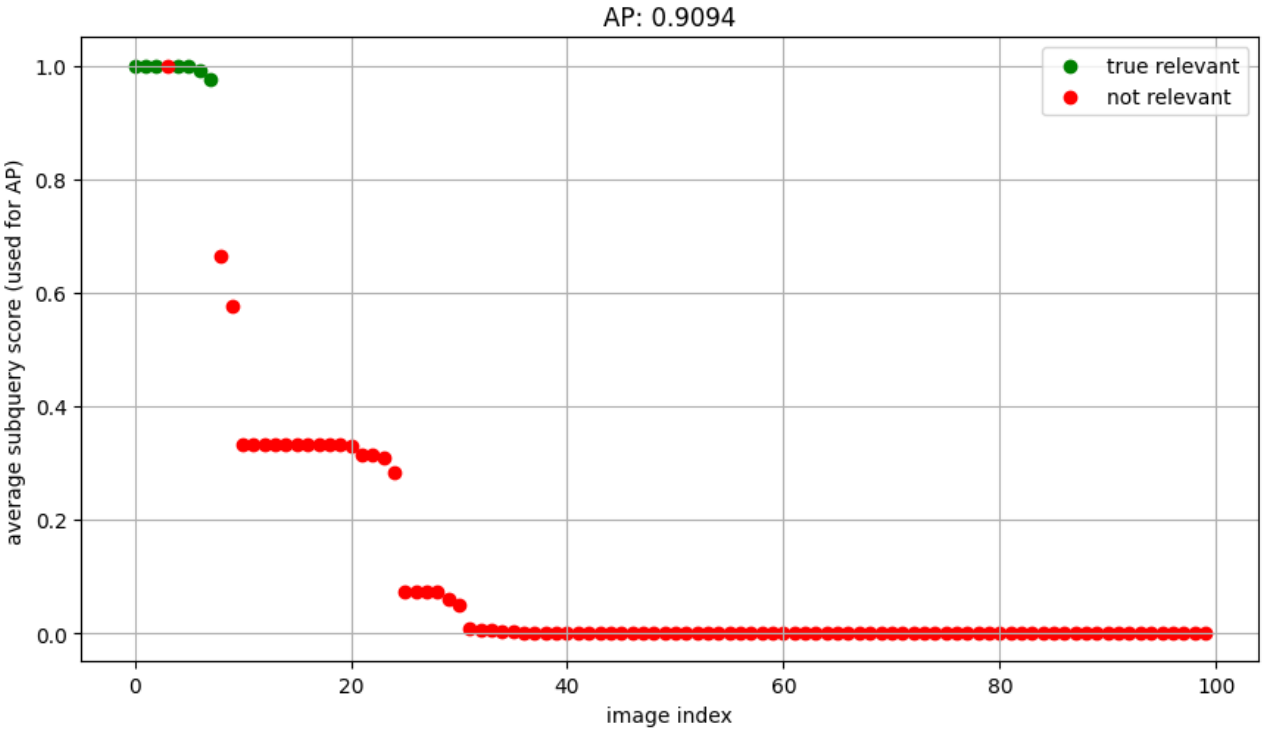


Figure 7. General approach to structuring the subquery generation prompt. The process begins by identifying the species or object of interest, followed by formulating visually verifiable questions that can be answered from images alone.



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