

## 1 A Implementation Details

2 QuARI’s transformer backbone is randomly initialized with 4 transformer layers. The query encoder  
3 and both the query and column decoders are two-layer MLPs with GeLU activation functions and  
4 layer normalization [1]. We train with a batch size of 320 and a contrastive temperature of 0.07. All  
5 backbone model embeddings are precomputed before training.

## 6 B Data Generation Prompt

7 For all datasets other than BioTrove [6], we use the provided natural language annotations as the  
8 text label. However, BioTrove does not provide natural language annotations outside of taxonomic  
9 and common-name identities. Therefore, we provide the species annotation along with the image to  
10 Qwen2.5-VL-7B-Instruct [5] with the following instruction:

11 “For the image shown, write one plain, human-sounding sentence that someone might type into an  
12 image search system to find this exact picture of a {species\_name}. Mention the main objects, their  
13 key attributes, and any distinctive action or setting. Keep it brief and objective, avoiding flowery  
14 descriptors unless they are essential to identify the scene. Output only this sentence.”

15 We collect these annotations on 500K images sampled from BioTrove to augment our training dataset  
16 with natural language descriptions of biodiversity-domain imagery.

## 17 C Broader Impacts

18 Improving retrieval systems to be both more accurate and more computationally efficient has broad  
19 positive implications, especially in domains where real-time or large-scale search is critical – such as  
20 recognizing where victims of human trafficking are photographed [4], monitoring biodiversity using  
21 camera trap images in ecological surveys [2], or identifying the spread of disinformation through  
22 manipulated visual media [3]. QuARI enables high-quality retrieval even with limited resources,  
23 making advanced search capabilities more accessible in a wider range of applications. We do not  
24 foresee unique negative societal impacts associated with QuARI beyond those that already exist with  
25 general-purpose image retrieval systems. Nevertheless, the broader implications of visual search  
26 technologies—including potential misuse in surveillance or disinformation—remain important areas  
27 for ongoing community oversight and ethical consideration.

## 28 References

- 29 [1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL  
30 <https://arxiv.org/abs/1607.06450>.
- 31 [2] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings*  
32 *of the European Conference on Computer Vision (ECCV)*, September 2018.
- 33 [3] Vinh Dang, Thanh-Son Nguyen, Minh-Triet Tran, and Duc-Tien Dang-Nguyen. Detecting  
34 misinformation in photos utilizing reverse image search. In *Proceedings of the 2024 International*  
35 *Conference on Multimedia Retrieval*, pages 1321–1323, 2024.
- 36 [4] Abby Stylianou, Hong Xuan, Maya Shende, Jonathan Brandt, Richard Souvenir, and Robert  
37 Pless. Hotels-50k: A global hotel recognition dataset. In *The AAAI Conference on Artificial*  
38 *Intelligence (AAAI)*, January 2019.
- 39 [5] Qwen Team. Qwen2.5-vl, January 2025. URL [https://qwenlm.github.io/blog/qwen2.](https://qwenlm.github.io/blog/qwen2.5-vl/)  
40 [5-vl/](https://qwenlm.github.io/blog/qwen2.5-vl/).
- 41 [6] Chih-Hsuan Yang, Benjamin Feuer, Talukder Jubery, Zi Deng, Andre Nakkab, Md Zahid Hasan,  
42 Shivani Chiranjeevi, Kelly Marshall, Nirmal Baishnab, Asheesh Singh, et al. Biotrove: A large  
43 curated image dataset enabling ai for biodiversity. *Advances in Neural Information Processing*  
44 *Systems*, 37:102101–102120, 2024.