
Composing Open-domain Vision with RAG for Ocean Monitoring and Conservation

Sepand Dyanatkar*, Angran Li, Alexander Dunate
OnDeck Fisheries AI

Abstract

Climate change’s destruction of marine biodiversity is threatening communities and economies around the world which rely on healthy oceans for their livelihoods. The challenge of applying computer vision to niche, real-world domains such as ocean conservation lies in the dynamic and diverse environments where traditional top-down learning struggle with long-tailed distributions, generalization, and domain transfer. Scalable species identification for ocean monitoring is particularly difficult due to the need to adapt models to new environments and identify rare or unseen species. To overcome these limitations, we propose leveraging bottom-up, open-domain learning frameworks as a resilient, scalable solution for image and video analysis in marine applications. Our preliminary demonstration uses pretrained vision-language models (VLMs) combined with retrieval-augmented generation (RAG) as grounding, leaving the door open for numerous architectural, training and engineering optimizations. We validate this approach through a preliminary application in classifying fish from video onboard fishing vessels, demonstrating impressive emergent retrieval and prediction capabilities without domain-specific training or knowledge of the task itself.

1 Introduction

Reliable and timely data is critical for addressing climate challenges driving global biodiversity collapse, and threatening 3 billion livelihoods that depend on healthy oceans [1]. Thousands of cameras are deployed around the world underwater, on drones, on fishing vessels, and via satellite to monitor marine life and inform critical management tasks. The scale of the data they produce is already impossible for humans to handle. In fisheries monitoring alone, thousands of hours of video are generated per boat, making its manual review crushingly slow, not scalable, and unaffordable. [2].

Automated processing of visual data brings key insights about marine activity at an unprecedented scale. Traditional vision approaches, such as CNNs and basic vision transformers [3, 4], require extensive retraining for new environments and perform poorly when faced with rare or unseen species. These top-down methods struggle with generalization, domain transfer, and handling long-tailed distributions, and are infeasible for scalable ocean applications. Generalizable species identification stands to revolutionize marine management by providing comprehensive data on marine life distribution and their responses to climate change - vital for management of sustainable fisheries, invasive species, carbon sequestration, and ecosystem health.

To overcome these challenges, we propose leveraging bottom-up, open-domain learning frameworks as a scalable and resilient solution, where knowledge about tuna species or oil spills are connected in a modular way, even after deployment. In this proposal, we show one accessible method for grounded, open-domain vision and highlight the promise of vision-language models (VLMs) combined with retrieval-augmented generation (RAG) [5–9]. This approach enhances generalization across diverse

*Corresponding author: sepand@ondeck-ai.com. Website: <https://ondeck.fish>

environments and enables accurate identification of unseen species by integrating external knowledge during inference.

2 Related Work

Traditional vision methods such as convolutional neural networks (CNNs), single-stage detectors, and vision transformers [4, 10, 11] have been the cornerstone of visual recognition. These top-down methods have shown reliable performance in closed-set, uni-domain climate applications, including environmental monitoring [12–14]. However, their effectiveness diminishes in generalized climate applications where data is often non-IID, and the distribution of objects and events is long-tailed or unexpected [15–18].

Recent advances in bottom-up, multi-modal approaches offer promising solutions to these challenges. Vision-language models (VLMs) based on CLIP and BLIP-2 [19, 20] learn representations of the world through contrastive learning and show emergent capabilities like detection, retrieval, question-answering and broader domain transfer. Note that this effectively bridges the modality gap between visual and textual data. These models excel in open-domain tasks such as visual question-answering (VQA) and retrieval [21, 22], which are crucial for climate monitoring and conservation [23].

The integration of grounding and retrieval-augmented generation (RAG) into these models further enhances their performance by allowing them to access and incorporate external knowledge during inference [5, 6, 24–26]. For instance, grounded CLIP [27] showcases emergent detection and identification capabilities that compete and surpass SOTA methods. Interestingly, Shen et al. [6] empirically indicates that biodiversity-related domains benefit most from transferable vision models injected with external knowledge (e.g. Flowers102 and OxfordPets). Grounding is key for open-ended VQA tasks, where external knowledge is required [21], but can also alleviate difficulties of class imbalance, domain shift and transfer, and interpretability. These advances make VLMs combined with RAG particularly promising for addressing the complex challenges posed by climate-related monitoring tasks.

Despite these advancements, there is still limited work that systematically composes, refines, and scales these approaches across diverse real-world applications. Our work addresses this gap by exploring the application of these methods in marine conservation, where the ability to generalize and adapt to new environments is critical for effective and scalable species identification and marine monitoring.

3 Method

Current vision research has produced powerful methods that can extrapolate and perform new tasks even without explicit training on them. While extremely generalizable, they cannot yet produce specialized outputs they have never seen in training, such as object classification to an unseen class. Thus grounding vision is critical as it provides access to external knowledge, and this is our focus for this preliminary work.

We outline a framework for building generalizable models adaptable to unseen domains and tasks, with a focus on climate impact, and marine monitoring applications. We integrate key components including bottom-up learning and grounding with retrieval-augmented generation (RAG), to address the challenges of scalability, adaptability, and robustness in real-world environments. For grounding and RAG, we show a minimal setup using similarity search, opening the door for complete grounding and pipelines in future work. Design enhancements, fine-tuning, and extra tools (such as prompt optimization, multi-query search, and large context) can greatly improve RAG-based approaches (see Appendix B).

We propose a methodology which uses an open-domain, bottom-up, and task-free model, specifically contrastive learning-style VLMs, combined with RAG to enhance model adaptability and performance across diverse and unseen domains. These models can easily transfer between domains, using modular RAG connections for grounding at any time and without model retraining. Similar to typical multi-modal retrieval-augmented transformer structures [5, 24, 28], our visual RAG pipeline has three components as shown in Fig. 1: a CLIP visual encoder to generate image embeddings, the knowledge base built with image embedding, and the backbone where we can evaluate various pre-trained

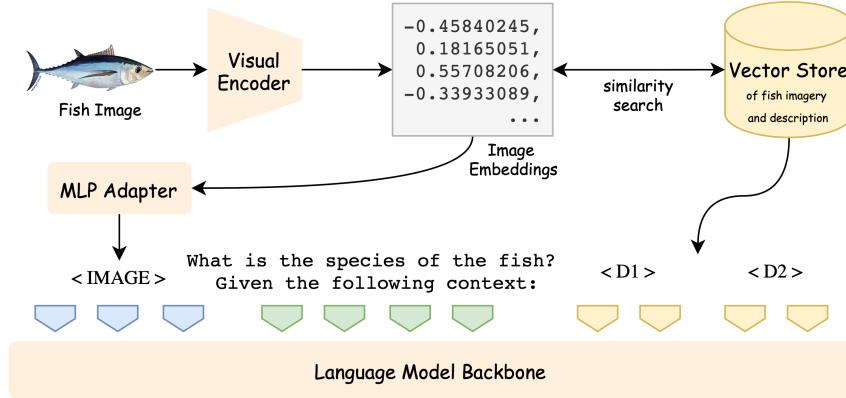


Figure 1: Architecture of visual RAG. The small pentagons with different colours represent tokens. They are concatenated as input into a language model to generate the final prediction.

language models. As this is preliminary work, we focus on demonstrating the value of information retrieval on visual classification and direct our investigation using an open dataset for fisheries monitoring [16]. For a comprehensive version of this method with attention to the entire pipeline, multiple domains, and multiple tasks, see Appendix B and future work.

3.1 Image-based Vector Store

In existing RAG-based vision work, vector stores use text-based keys [25, 26, 29]. Our work is the first to our knowledge to build the knowledge base with an image embedding key. It is motivated by several reasons: a) Poor image quality (Fig. 2) causes direct similarity search between these images’ features and text to result in noise and confuses the language model; b) The image embedding database leverages the limited amount of labeled data, helping constrain knowledge retrieval; c) The query itself is static (e.g. object classification), hence adaptive retrieval according to user prompt is unnecessary.

We produce image embeddings for a small set of reference species using the CLIP encoder and store them as the key in a vector database. The descriptions of different species can then be retrieved by their corresponding image embedding and used as augmented context for answer generation in the final step. Specifically, we produce embeddings of images in the Fishnet validation set using the CLIP encoder and store them as the key in our vector database.

3.2 Pre-trained Multi-modal LLM

The LLaVA [30] family is widely used as the backbone for multi-modal downstream tasks. We modify the CLIP [19] model as visual encoder to generate the query for the similarity search, and use the pre-trained LLaVA 1.5 weights for performance evaluation. To perform a new *set* of tasks (while still) in an emergent manner, VLMs greatly benefit from instruction fine-tuning [7], which can be supported with LoRA[31]. We leave this for future work (see Appendix B) building on this proposal.

4 Preliminary Evaluation

As was selected for motivation and demonstration in the methods section, we investigate visual RAG on the sample visual task of classification for real-world climate applications.

We implemented our proposed method using the exact minimal architecture shown of Fig. 1. As our dataset, we use Kay and Merrifield [16]’s Fishnet dataset, version 1.0.0, since it is the largest public dataset of on-deck fisheries activity, and represents a challenge in many of the dimensions discussed in previous sections (data distribution, domain shifts, unseen knowledge).

For visual RAG, we evaluated final prediction performance as this is most comparable to baselines for a classification task (see Table 1). We also test visual RAG’s intermediate accuracy at the information

Table 1: Classification accuracy of baseline vs. our VLM-RAG approach on 5 categories. We measure both performance of final prediction (single answer response) and intermediate RAG retrieval.

Method	Top-1	Accuracy	
		Top-2	Top-3
InceptionV3 (Baseline)	0.7501	0.8312	0.9408
VLM-RAG (Ours, Final Prediction)	0.8403	N/A (single answer)	N/A (single answer)
VLM-RAG (Ours, RAG Retrieval)	0.8684	0.9527	0.9781

retrieval step, investigating the performance of the encoder vector search separately. Completing our preliminary ablation, we also tested the pipeline without any grounding (see Fig. 3). As our baseline, we used InceptionV3 pretrained on ImageNet with all the same parameters as provided in Kay and Merrifield [16] using a single NVIDIA A100.

For intuition on the task difficulty and influence of retrieved descriptions, Fig. 2 shows typical input image, prompt, and output, using low resolution and partially occluded input, both detrimental for species classification. Without retrieval visually shows our ablation where the category list is explicitly given to the model.

Even with no re-ranking or other optimizations in this preliminary implementation, Table 1 shows retrieval performance outperforming both baseline and final prediction. We predict that by advancing the methods outlined in Section 3 with steps described in Appendix B and high-quality embedded descriptions and sample images, the final prediction will outperform RAG retrieval while both see an additional increase.

In the appendix, we present top-k accuracy of the retrieval process (Fig. 4), capturing whether the language model has enough information for accurate prediction. Figure 3 then shows the precision and recall of the final prediction under different RAG settings. Additionally, we give a visualisation of our embedding space (Fig. 5), to emphasize the task difficulty and the necessity of using visual RAG on on-deck fish images.

5 Conclusion and Pathways to Climate Impact

Our results demonstrate impressive retrieval and prediction capabilities, without any task or domain-specific training, highlighting the potential of bottom-up learning models to advance scalable marine monitoring. Unblocked from the need for expensive domain adaptation, our continued collaboration with fisheries and marine conservation partners will enable faster and more accessible deployments of marine life monitoring for significantly more informed responses to changing climates.



With retrieval:

Q: What is the species of the fish?

A: The fish in the image is a tuna, as indicated by its torpedo-shaped body, small dorsal and pectoral fins, and metallic blue coloration.

Without retrieval, choices provided:

Q: What is the species of the fish?

A: The fish in the image is a Mahi Mahi.

Figure 2: Example input image and QA with RAG retrieved description (not shown in figure) and without RAG (category list provided but not shown). Images are often low resolution and partly occluded.

Acknowledgements

We would like to thank the entire OnDeck Fisheries AI team for supporting our research and their hard work to translate research into real solutions, accessible and affordable around the world. Thank you to Prof. Graham Taylor’s research group in 2023 for conversations about combining non-text modalities with RAG for biodiversity science which encouraged us to conduct our initial literature review and begin this work.

References

- [1] *Goal 14 : Life below water*. Joint SDG Fund | Goal 14: Life below water. URL: <https://jointsdgfund.org/sustainable-development-goals/goal-14-life-below-water> (cit. on p. 1).
- [2] Kate Wing and Benjamin Woodward. “Advancing artificial intelligence in fisheries requires novel cross-sector collaborations”. In: *ICES Journal of Marine Science* (Aug. 28, 2024). Ed. by Howard Browman, fsae118. ISSN: 1054-3139, 1095-9289. URL: <https://academic.oup.com/icesjms/advance-article/doi/10.1093/icesjms/fsae118/7742959> (cit. on p. 1).
- [3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc., 2012. URL: https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html (cit. on p. 1).
- [4] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: International Conference on Learning Representations. Oct. 2, 2020. URL: <https://openreview.net/forum?id=YicbFdNTTy> (cit. on pp. 1, 2).
- [5] Davide Caffagni et al. *Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs*. May 22, 2024. arXiv: 2404.15406[cs]. URL: <http://arxiv.org/abs/2404.15406> (cit. on pp. 1, 2).
- [6] Sheng Shen et al. “K-LITE: Learning Transferable Visual Models with External Knowledge”. In: *Advances in Neural Information Processing Systems* 35 (Dec. 6, 2022), pp. 15558–15573. URL: https://papers.nips.cc/paper_files/paper/2022/hash/63fef0802863f47775c3563e18cbb17-Abstract-Conference.html (cit. on pp. 1, 2).
- [7] Zalan Fabian et al. “Knowledge Augmented Instruction Tuning for Zero-shot Animal Species Recognition”. In: NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following. Nov. 26, 2023. URL: <https://openreview.net/forum?id=0QHckRYbpT> (cit. on pp. 1, 3).
- [8] Kenneth Marino et al. “KRISP: Integrating Implicit and Symbolic Knowledge for Open-Domain Knowledge-Based VQA”. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, June 2021, pp. 14106–14116. ISBN: 978-1-66544-509-2. URL: <https://ieeexplore.ieee.org/document/9577534/> (cit. on p. 1).
- [9] Jialin Wu et al. *Multi-Modal Answer Validation for Knowledge-Based VQA*. Dec. 13, 2021. arXiv: 2103.12248[cs]. URL: <http://arxiv.org/abs/2103.12248> (cit. on p. 1).
- [10] Kaiming He et al. *Deep Residual Learning for Image Recognition*. Dec. 10, 2015. arXiv: 1512.03385[cs]. URL: <http://arxiv.org/abs/1512.03385> (cit. on p. 2).
- [11] Joseph Redmon et al. “You Only Look Once: Unified, Real-Time Object Detection”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, June 2016, pp. 779–788. ISBN: 978-1-4673-8851-1. URL: <http://ieeexplore.ieee.org/document/7780460/> (cit. on p. 2).
- [12] Mohammad Sadegh Norouzzadeh et al. “Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning”. In: *Proceedings of the National Academy of Sciences* 115.25 (June 19, 2018). Publisher: Proceedings of the National Academy of Sciences, E5716–E5725. URL: <https://www.pnas.org/doi/full/10.1073/pnas.1719367115> (cit. on p. 2).

- [13] Mohammad Sadegh Norouzzadeh et al. *A deep active learning system for species identification and counting in camera trap images*. Oct. 21, 2019. arXiv: 1910.09716[cs, eess, stat]. URL: <http://arxiv.org/abs/1910.09716> (cit. on p. 2).
- [14] William I. Atlas et al. “Wild salmon enumeration and monitoring using deep learning empowered detection and tracking”. In: *Frontiers in Marine Science* 10 (2023). ISSN: 2296-7745. URL: <https://www.frontiersin.org/articles/10.3389/fmars.2023.1200408> (cit. on p. 2).
- [15] Grant Van Horn et al. “The iNaturalist Species Classification and Detection Dataset”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT: IEEE, June 2018, pp. 8769–8778. ISBN: 978-1-5386-6420-9. URL: <https://ieeexplore.ieee.org/document/8579012/> (cit. on p. 2).
- [16] Justin Kay and Matt Merrifield. *The Fishnet Open Images Database: A Dataset for Fish Detection and Fine-Grained Categorization in Fisheries*. June 16, 2021. arXiv: 2106.09178[cs]. URL: <http://arxiv.org/abs/2106.09178> (cit. on pp. 2–4).
- [17] Kakani Katija et al. “FathomNet: A global image database for enabling artificial intelligence in the ocean”. In: *Scientific Reports* 12.1 (Sept. 23, 2022). Number: 1 Publisher: Nature Publishing Group, p. 15914. ISSN: 2045-2322. URL: <https://www.nature.com/articles/s41598-022-19939-2> (cit. on p. 2).
- [18] Samuel Stevens 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 (CVPR)*. 2024, pp. 19412–19424 (cit. on p. 2).
- [19] Alec Radford et al. “Learning Transferable Visual Models From Natural Language Supervision”. In: *Proceedings of the 38th International Conference on Machine Learning*. International Conference on Machine Learning. ISSN: 2640-3498. PMLR, July 1, 2021, pp. 8748–8763. URL: <https://proceedings.mlr.press/v139/radford21a.html> (cit. on pp. 2, 3).
- [20] Junnan Li et al. *BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models*. June 15, 2023. arXiv: 2301.12597[cs]. URL: <http://arxiv.org/abs/2301.12597> (cit. on p. 2).
- [21] Kenneth Marino et al. “OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019). Conference Name: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) ISBN: 9781728132938 Place: Long Beach, CA, USA Publisher: IEEE, pp. 3190–3199. URL: <https://ieeexplore.ieee.org/document/8953725/> (cit. on p. 2).
- [22] Jun Xu et al. “MSR-VTT: A Large Video Description Dataset for Bridging Video and Language”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, June 2016, pp. 5288–5296. ISBN: 978-1-4673-8851-1. URL: <http://ieeexplore.ieee.org/document/7780940/> (cit. on p. 2).
- [23] Zhongqi Miao et al. “New frontiers in AI for biodiversity research and conservation with multimodal language models”. In: (Aug. 1, 2024). Publisher: EcoEvoRxiv. URL: <https://ecoevorxiv.org/repository/view/7477/#!> (cit. on p. 2).
- [24] Sara Sarto et al. *Retrieval-Augmented Transformer for Image Captioning*. Aug. 22, 2022. arXiv: 2207.13162[cs]. URL: <http://arxiv.org/abs/2207.13162> (cit. on p. 2).
- [25] Zhengbao Jiang et al. *Active Retrieval Augmented Generation*. Oct. 21, 2023. arXiv: 2305.06983[cs]. URL: <http://arxiv.org/abs/2305.06983> (cit. on pp. 2, 3).
- [26] Kelvin Guu et al. *REALM: Retrieval-Augmented Language Model Pre-Training*. Feb. 10, 2020. arXiv: 2002.08909[cs]. URL: <http://arxiv.org/abs/2002.08909> (cit. on pp. 2, 3).
- [27] Liunian Harold Li et al. “Grounded Language-Image Pre-training”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, June 2022, pp. 10955–10965. ISBN: 978-1-66546-946-3. URL: <https://ieeexplore.ieee.org/document/9879567/> (cit. on p. 2).
- [28] Rita Ramos et al. *SmallCap: Lightweight Image Captioning Prompted with Retrieval Augmentation*. Mar. 28, 2023. arXiv: 2209.15323[cs]. URL: <http://arxiv.org/abs/2209.15323> (cit. on p. 2).

- [29] Ziniu Hu et al. *REVEAL: Retrieval-Augmented Visual-Language Pre-Training with Multi-Source Multimodal Knowledge Memory*. Apr. 3, 2023. arXiv: 2212.05221[cs]. URL: <http://arxiv.org/abs/2212.05221> (cit. on p. 3).
- [30] Haotian Liu et al. *Visual Instruction Tuning*. Dec. 11, 2023. arXiv: 2304.08485[cs]. URL: <http://arxiv.org/abs/2304.08485> (cit. on p. 3).
- [31] Edward J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. Oct. 16, 2021. arXiv: 2106.09685[cs]. URL: <http://arxiv.org/abs/2106.09685> (cit. on p. 3).

A Additional Evaluation Details

A.1 Precision and Recall by Category w/wo Retrieved Description

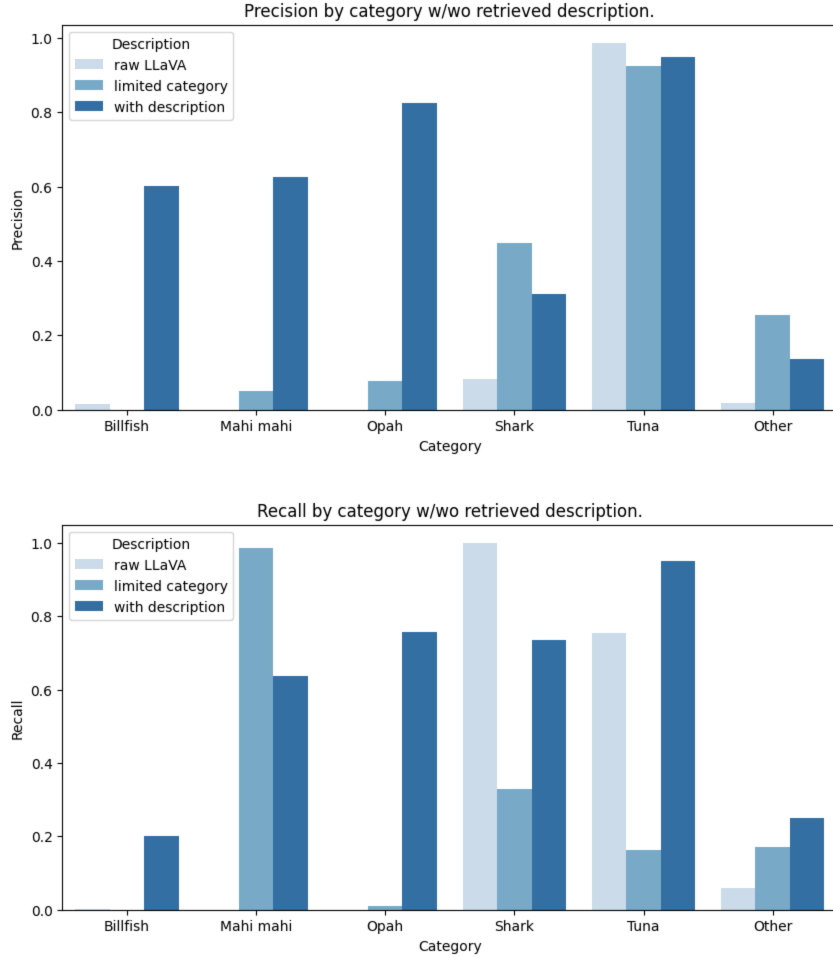


Figure 3: Precision and recall by category in different experiment settings.

Figure 3 shows LLaVA final prediction precision and recall by category. The experiment is conducted in three different settings:

- Raw LLaVA without any categories provided or information retrieved from the database.
- LLaVA with category provided and limited to Billfish, Mahi mahi, Opah, Shark, Tuna, and Other.
- LLaVA with retrieved description. Here the category limit is implicitly included by description retrieval.

A.2 Top-k Accuracy

Figure 4 shows the top-k accuracy of the retrieval process. That is, we are measuring how often the top k retrieved descriptions contain the correct category or species. Higher accuracy of the retrieval should typically mean the model has higher probability to make the correct prediction in the generation step, since it is provided with higher accuracy external knowledge.

The experiments are conducted in two different granularities: category and species, where one category might contain several species that are hard to distinguish. For example, category Tuna contains species Albacore, Yellowfin tuna, Skipjack tuna, Bigeye tuna, and Tuna (which the dataset grouped due to ambiguity).

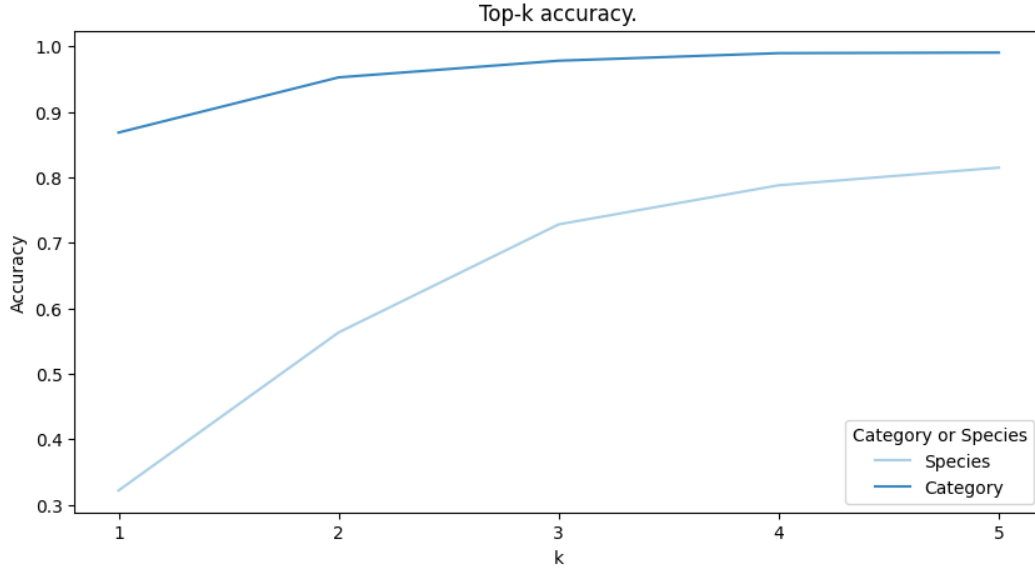


Figure 4: Top-k accuracy for the RAG retrieval process.

A.3 Embedding Space Visualization

Figure 5 visualizes the image embedding of our vector store and test set samples, using PCA for dimension reduction. The Figures show:

- The sample feature distribution between two data sets are generally similar to each other. For example, orange samples (Billfish) are in top-right and blue samples (Tuna) are in bottom left.
- However, it is still fair to say that there are many differences between the two distributions. For example, the orange samples (Billfish) in the test set are closer to other categories than the corresponding class in the vector store.
- The embedding visualization also demonstrates that features from different categories in one data set are fairly mixed up, making it hard to identify the fish species.

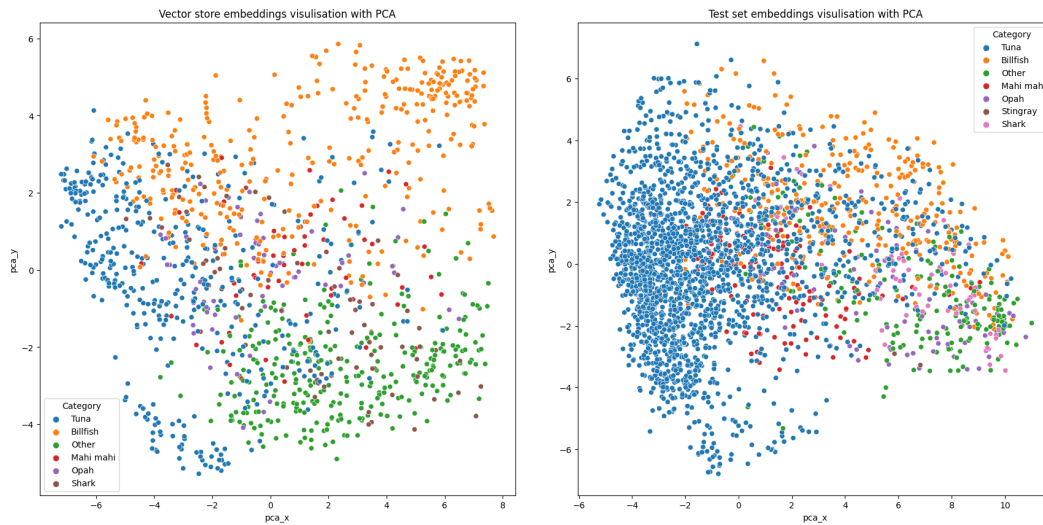


Figure 5: Image embedding 2D visualization of vector store and test set.

B Discussion and Future Work

The promise of grounded visual reasoning is quite significant. This is even more pronounced with the current momentum towards universal reasoners based on projections of scaling laws for vision models. Grounding these highly generalizable methods can allow us to reach human capacity for analyzing imagery, where a model will look up information it does not have just as a human would. We also emphasize that this work does not explore the "reasoning design", in our case being the choice of (static) pipeline steps and leading into a final VLM-as-predictor with a static query. For example, introducing guided reasoning would allow the architecture to be dynamically adjusted, potentially addressing multi-task capability, and recursive reasoning and attending to harder queries.

Visual reasoning with RAG is generally underexplored. We anticipate and encourage numerous creative and powerful solutions to build on our proposed methodology.

A brief list of dimensions for future work which we anticipate and recommend:

- VLM selection and fine-tuning for the task.
- RAG-specific improvements (examples in Fig 6):
 - Multi-query searching
 - Improved search algorithms
 - Prompt optimization
 - Re-ranking results
 - Optimization of the vector store storage for multiple modalities (going beyond just image and text).
 - Increased number of results from vector store.
- Extracting various streams of information from the imagery (related to multi-query searching). E.g., one stream of nearby background or context, one stream of information about relative sizes of everything in the scene, one stream of just the object in question, etc.
- Optimizing for hierarchical step-by-step reasoning, by descending in dimensions of specificity and complexity.
- Runtime/In-situ addition & removal of RAG databases
- Extensive evaluation against data with quantifiable and varying human agreement on labels.

With all of these enhancements, we also expect a complete ablation to verify the enhancements.

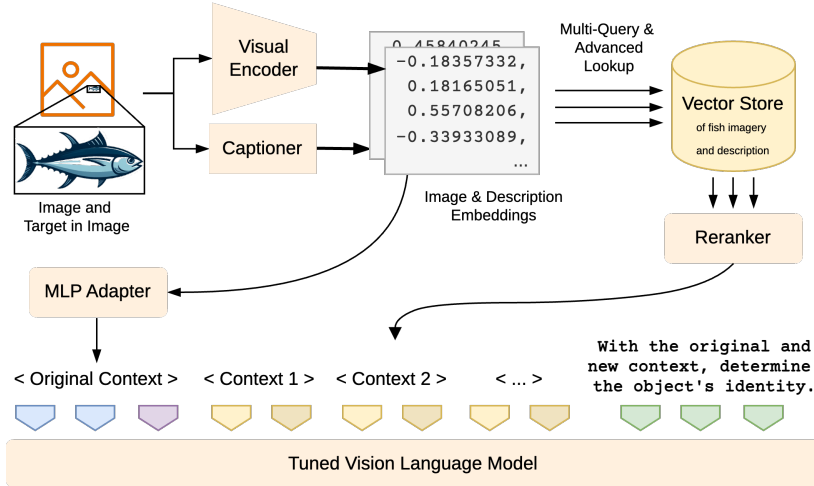


Figure 6: Next iteration of our proposed architecture for visual RAG in classification.