RealBirdID: Benchmarking Bird Species Identification in the Era of MLLMs

Supplementary Material

A. Different Abstention Thresholding Criteria

As mentioned in the encoder abstention metrics section, the choice of max probability to create a threshold for abstention tradeoff is not obvious. In this section, we experiment with two different methods for creating a decision threshold: (1) entropy, (2) max probability, and (3) probability difference.

The difference between criteria is small. In Fig. 1 we show the abstention tradeoff for various models using the three criteria. We find that in all cases, models hover around random prediction with the best results being achieved by probability difference. Similarly, in Fig. 2 we see that varying the criteria for these models does not significantly affect classification performance, with the best two criteria being max probability and probability difference.

Max probability is less correlated with choice count. A large portion of our experiments involve subsetting the possible list of species for a given observation by SINR. Naturally, this introduces a discrepancy in the amount of choices depending on the observation. We hypothesize that this caused an effect on the average entropy for each choice count, depicted in Fig. 3. There, we see that entropy increases as a steady function of choice count. However, we see that max probability does not have a clear correlation with choice count. This means that using max probability as a criteria less susceptible to simply predicting images with higher species counts as confusing in the abstention tradeoff.

B. Additional Experiment Details

Querying Proprietary APIs. For the proprietary MLLMs GPT-5 and Gemini-2.5 Pro, we use the single-turn API calls using the same prompt `"What is the species of this bird?"` and the input image. No system prompts, auxiliary instructions, or tool calls were used. Images were transmitted as base64-encoded JPEGs according to the providers' multimodal specifications. We set the reasoning-effort parameter to its minimal value (`minimal` for both models) and kept all other generation parameters at their documented defaults, including the default sampling temperature and no maximum output token limit. In this configuration, GPT-5 typically produced responses with negligible explicit reasoning tokens, whereas Gemini-2.5 Pro often emitted a short chain-of-thought before the final answer (87 tokens on average), resulting in a modest but consistent reasoning-token overhead. Across the answerable (31,885) and unanswerable (3,253) subsets, we issued 35,138 multimodal API calls per model.

BioCLIP-2 leads the pack in terms of classification performance. We find the best performing encoder to be BioCLIP-2, having the highest species classification abilities ($AUC =$ **.567**), genus classification abilities ($AUC =$ **.934**), and information gain ($AUC =$ **.845**). In particular, we find the out-of-the-box accuracies on the species (3561-way) and genus (248-way) levels to be **41%** and **76%** respectively, very strong performance for such a large multi-way task.

Training data leakage. To assess whether these results could be attributed to training data leakage, we conducted a comprehensive overlap analysis between BioCLIP-2's training corpus (TreeOfLife-200M) and our test sets. Each test image is associated with an iNaturalist observation ID, observation URL, and photo URL, while each TOL image contains a single `source_url` field. We canonicalized all identifiers on both sides, mapping each test-set identifier and each TOL `source_url` to a normalized photo ID, observation ID, or URL, and compared these canonical keys for exact and near-exact matches. Here, "near-exact" refers to different URL forms of the same underlying media (e.g., different resolutions or hostnames that map to the same iNaturalist photo or observation ID). Across the full 213.9M TOL images, we identified overlaps for 56.5% of samples in the answerable test set ($18{,}007/31{,}885$ images) and 1.1% in the unanswerable set ($36/3{,}253$ images). This substantial overlap suggests possible training data leakage and may partially explain BioCLIP-2's high performance. These findings highlight the need for future evaluations on fully de-duplicated benchmarks to more rigorously assess out-of-distribution generalization.

C. Additional Details

Ethical and Licensing Considerations. This release is intended for validation and analysis only. We used whatever media licenses were returned by the API; in the public release, we will (i) filter to permitted licenses (e.g., CC-BY/CC-BY-NC) and (ii) include clear provenance to original observations, respecting iNaturalist's terms and any location obscuration for sensitive taxa. Because we do not provide training splits, we also avoid any leakage between unanswerable and answerable resources by not reusing the exact observation images across sections.

Limitations. Firstly, the unanswerable dataset is inherently imbalanced: certain genera are overrepresented due to uneven observation rates, and some species pairs are more prone to visual ambiguity than others. This imbalance may influence both model behavior and evaluation metrics. Second, the labeling of unanswerable examples depends on expert judgment. For example, experts may disagree on what constitutes an "obstructed view" or whether a particular image lacks sufficient evidence for identification.

Future work could extend this framework to multimodal settings, incorporating optional modalities such as sound recordings, temporal context, and multiple observations. Another direction is improving the abstention calibration, particularly for multimodal MLLMs, which tend to overcommit despite uncertainty.

Although this work aims to promote responsible deployment through abstention-aware modeling, potential negative societal impacts should be considered. Miscalibrated abstention or overconfident misclassification may undermine public trust in these tools, especially when integrated into citizen-science platforms. Overreliance on model outputs could also discourage human expertise or misinform conservation decisions if abstention signals are misunderstood. To mitigate these risks, future iterations of the benchmark and accompanying systems should emphasize transparency, interpretability, and human-in-the-loop evaluation.

D. Organization of Remaining Figures.

1. Spacial distribution of observations seen within the answerable and unanswerable data. Fig. 4

2. Distribution of species per observation and per taxon. Fig. 5

3. MANIQA Distribution of RealBirdID vs. CUB200 Fig. 6

4. RealBirdID iNaturalist observation dates. Fig. 7

5. Table of most common genera occuring in the unanswerable data. Tab. 1

6. Classification results using class-averaging on the answerable and unanswerable sets. Fig. 8

7. Detailed classification curve tradeoffs and compiled range map information. Fig. 9, Fig. 10, Fig. 11

8. 100 examples randomly chosen from the answerable set. Fig. 12

9. 30 unanswerable examples which have abstention reason of "angle/occlusion." Fig. 13

10. 30 unanswerable examples which have abstention reason of "vocalization." Fig. 14

11. 30 unanswerable examples which have abstention reason of "angle/occlusion." Fig. 15

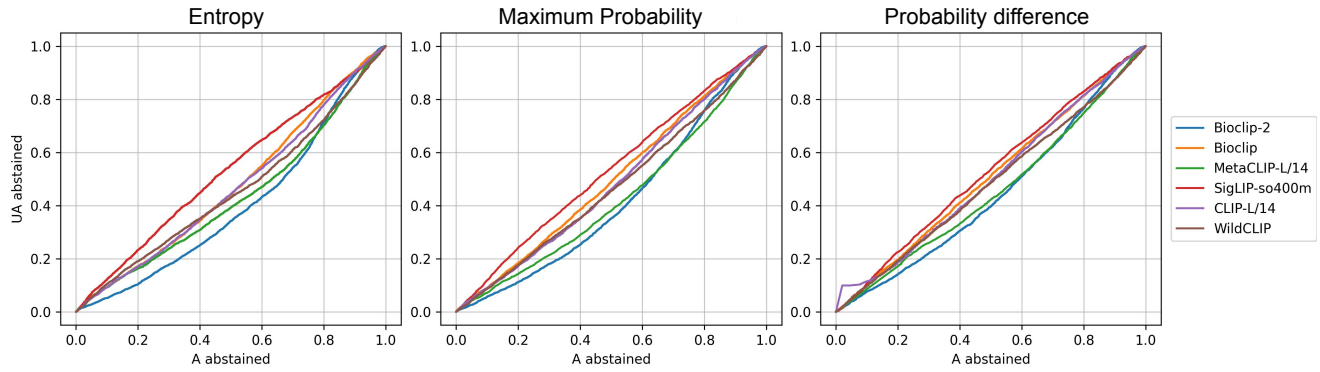


Figure 1. **Abstention tradeoff curves for different abstention criteria.**

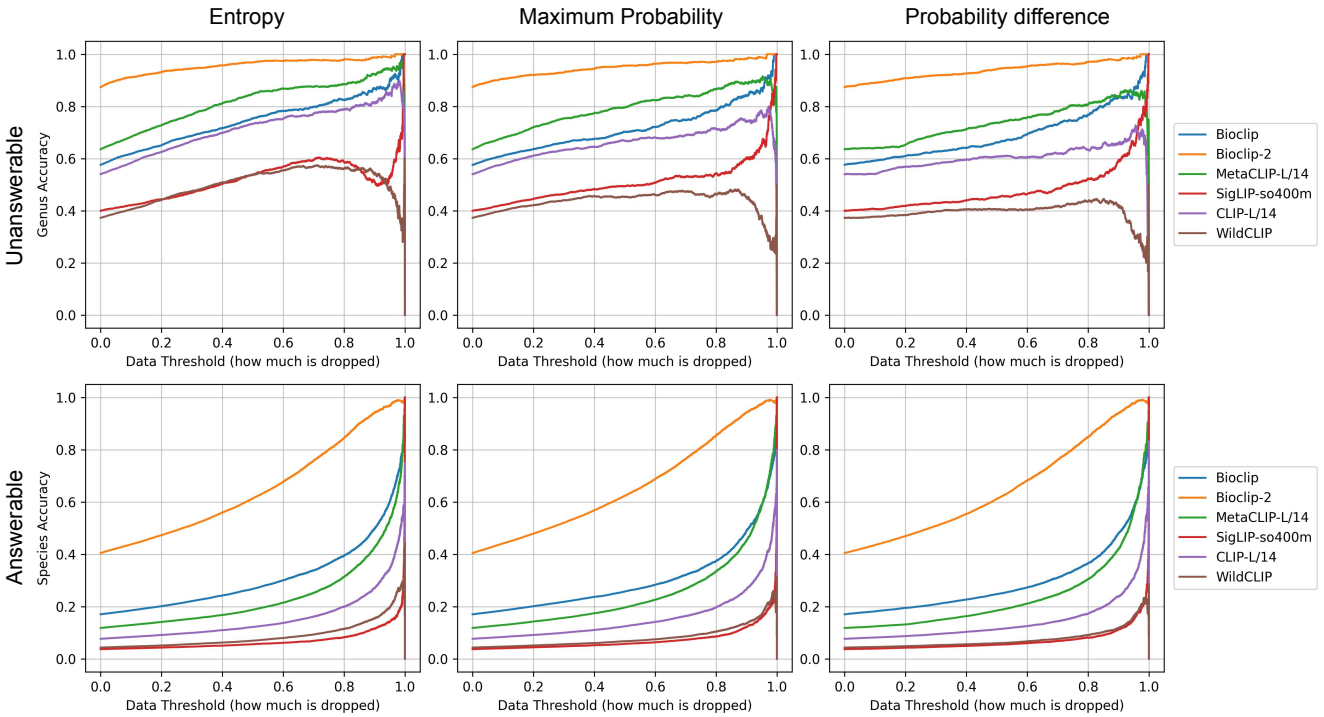


Figure 2. **Accuracy vs thresholding for different abstention criteria.**

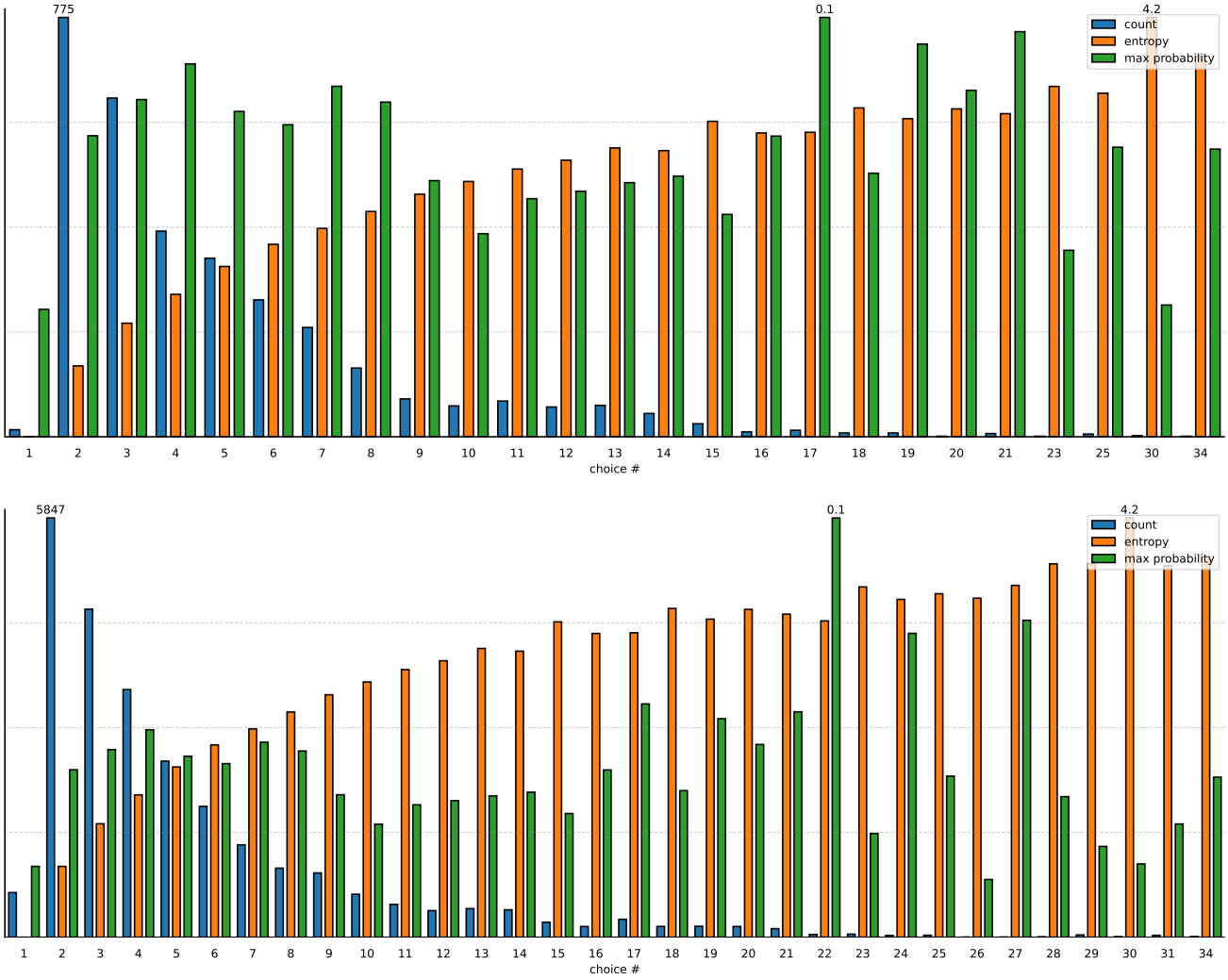


Figure 3. **Frequency, probability entropy, and max probability for various choice counts in the Range Info Species Subset data.** For the Range Info data, the examples are split up by the amount of choices that the model must choose between. For each choice count, the frequency, entropy, and max probability are shown. Bars are normalized by series. (**Top**) depicts the answerable set, whereas (**bottom**) depicts the unanswerable set. Both are run with Qwen-2.5VL-7B.

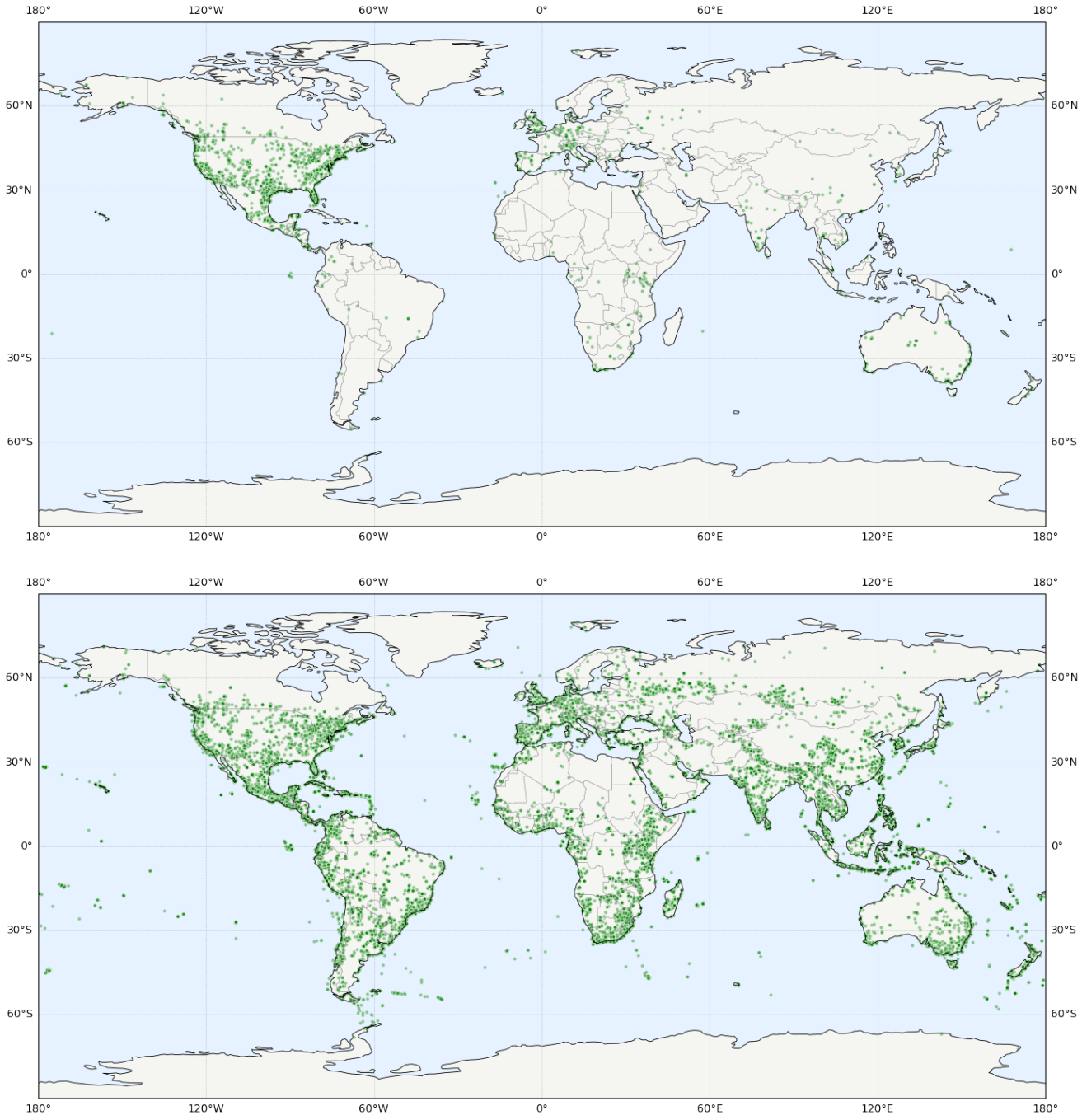


Figure 4. **Location info of the Unanswerable (top) and Answerable (bottom) sets.**

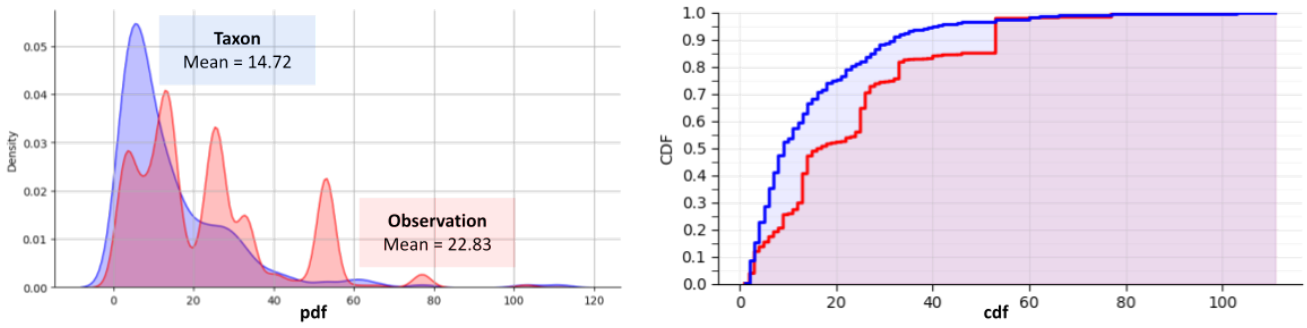


Figure 5. **Distribution of species per observation and per taxon.**

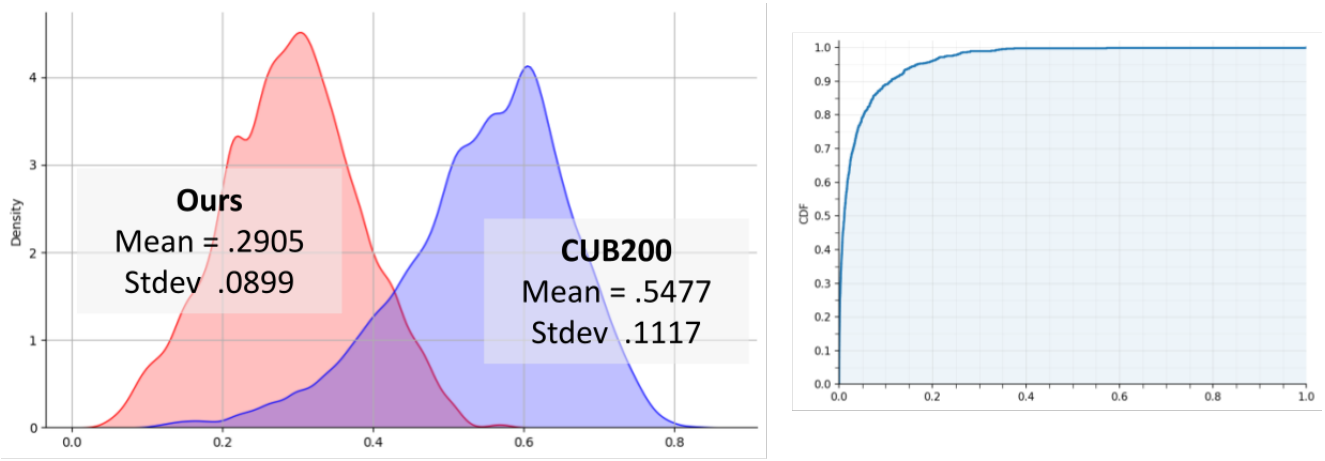


Figure 6. **MANIQA distributions of RealBirdID (ours) vs. CUB200.** On average, we find that RealBirdID has lower MANIQA scores.

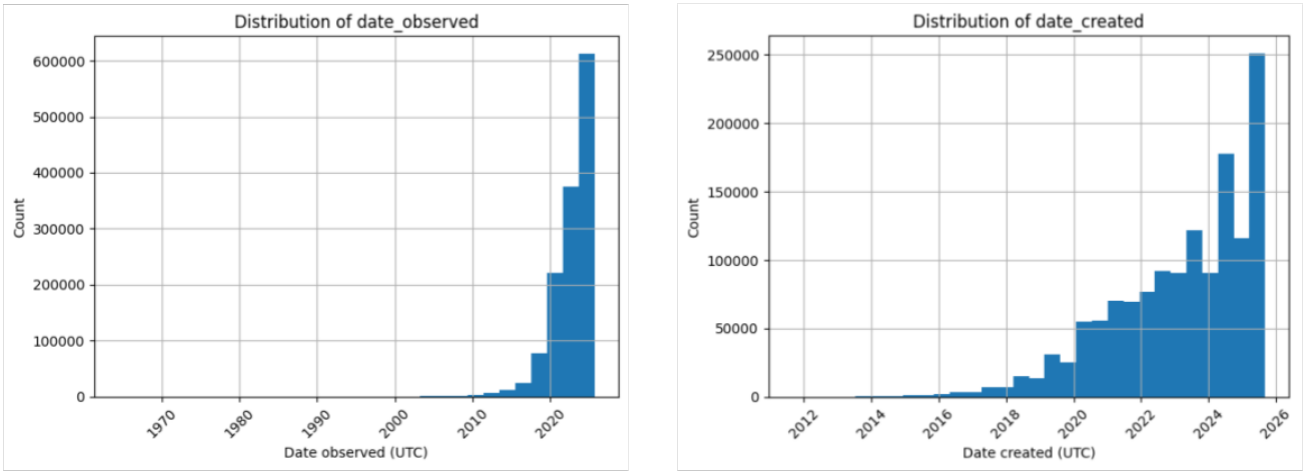
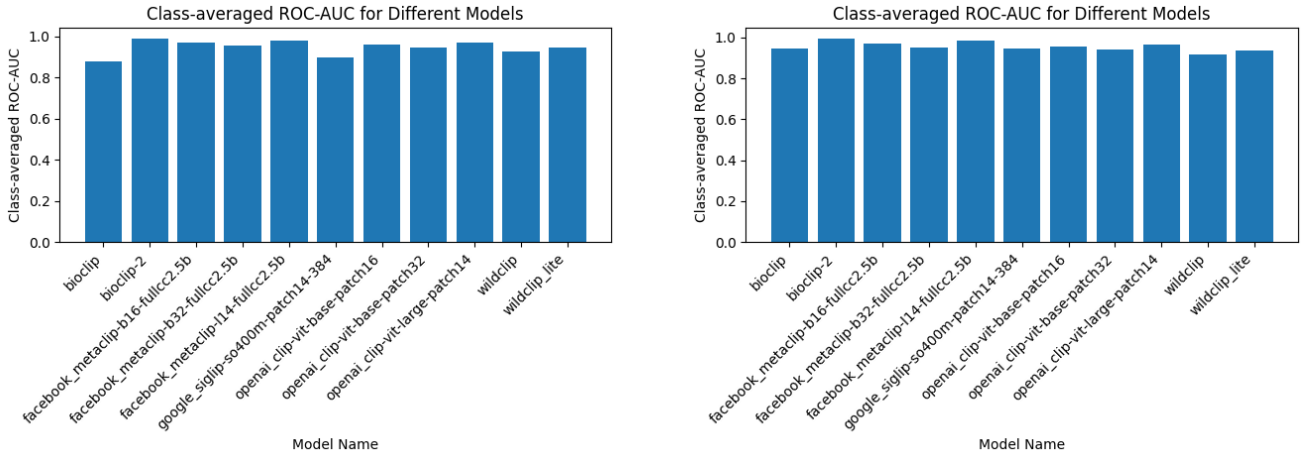


Figure 7. **Observation date distribution by observed and created timestamps.**

genus	count	freq	cdf	species
Crows and Ravens	291	0.1261	0.1261	53
Large White-headed Gulls	198	0.0858	0.2119	25
Kingbirds	179	0.0776	0.2894	13
Empidonax Flycatchers	132	0.0572	0.3466	14
Mallards, Pintails, and Allies	125	0.0542	0.4008	33
Dryobates Woodpeckers	116	0.0503	0.4510	26
Rufous, Allen's, and Allied Hummingbirds	53	0.0230	0.4740	9
Buteos	52	0.0225	0.4965	27
Dowitchers	47	0.0204	0.5169	3
Calidris Sandpipers	39	0.0169	0.5338	24
Scaups, Pochards, and Allies	39	0.0169	0.5507	12
True Swans	37	0.0160	0.5667	9
Leaf Warblers	35	0.0152	0.5819	77
Yellow-breasted Meadowlarks	32	0.0139	0.5958	3
Shanks, Tattlers, and Allies	27	0.0117	0.6075	13
Chickadees and Allies	27	0.0117	0.6192	15
Ruby-throated and Black-chinned Hummingbirds	25	0.0108	0.6300	2
Plegadis Ibises	24	0.0104	0.6404	3
Typical Falcons	23	0.0100	0.6503	40
American Cormorants	20	0.0087	0.6590	3
Brown Thrushes and Nightingale-Thrushes	20	0.0087	0.6677	13
Great Herons and Egrets	19	0.0082	0.6759	17
Western and Clark's Grebes	19	0.0082	0.6841	2
Setophaga Warblers	19	0.0082	0.6924	34
Yellow-tailed and White-tailed Black Cockatoos	18	0.0078	0.7002	3
Vireos	17	0.0074	0.7075	33
True Sparrows	15	0.0065	0.7140	28
...				
Trillers and Allies	1	0.0004	0.9991	20
Typical White-eyes	1	0.0004	0.9996	111
Locustellid Bush Warblers and Allies	1	0.0004	1.0000	23

Table 1. **Most common generas occuring in the unanswerable data.**



Binary Classification Results on the Answerable Subset (Left)

Binary Classification Results on the Unanswerable Subset (Right)

Figure 8. **Classification results using class-averaging on the answerable and unanswerable sets.**

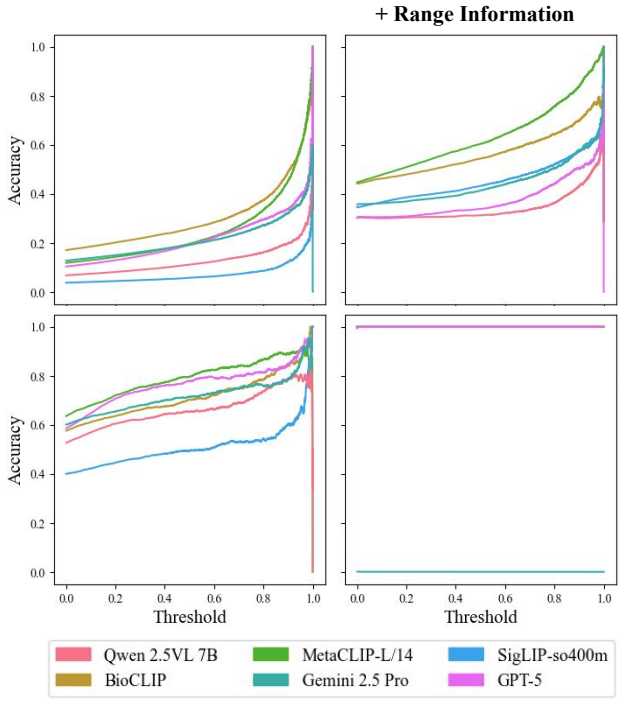


Figure 9. **Classification performance of CLIP-based models and popular MLLMs on the species level of answerables (A) and the genus level of unanswerables (UA).** For various CLIP encoders, accuracies at percentile-based max probablities are plotted when sweeping over percent of data thresholded. For the answerable set the species label is used to compute accuracy **(top)** whereas for the unanswerable set, the genus label is used **(bottom)**. On the **(right)** we observe the effect of using *species range maps* to constrict the choice set. Note that the genus accuracy for encoders is *not an error*: subsetting species list using SINR location info achieves perfect genus performance.

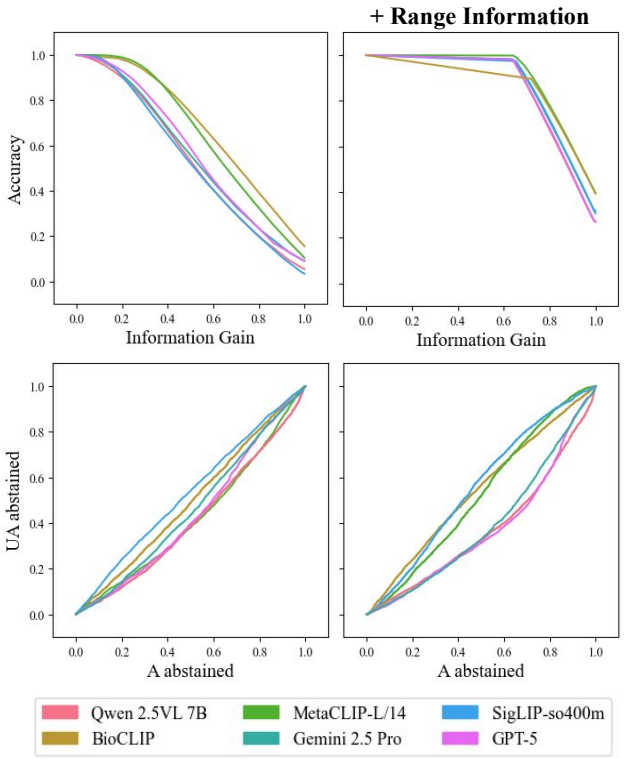


Figure 10. **Abstention calibration and entropy–threshold selective classification.** For each model we sweep an entropy threshold on the flat species-level softmax and plot the fraction of unanswerable (UA) examples abstained on against the fraction of answerable (A) examples abstained **(bottom)**. We combine the UA / A performance into a unified classification metric using Information Gain vs. Accuracy as proposed by DARTS (**top**).

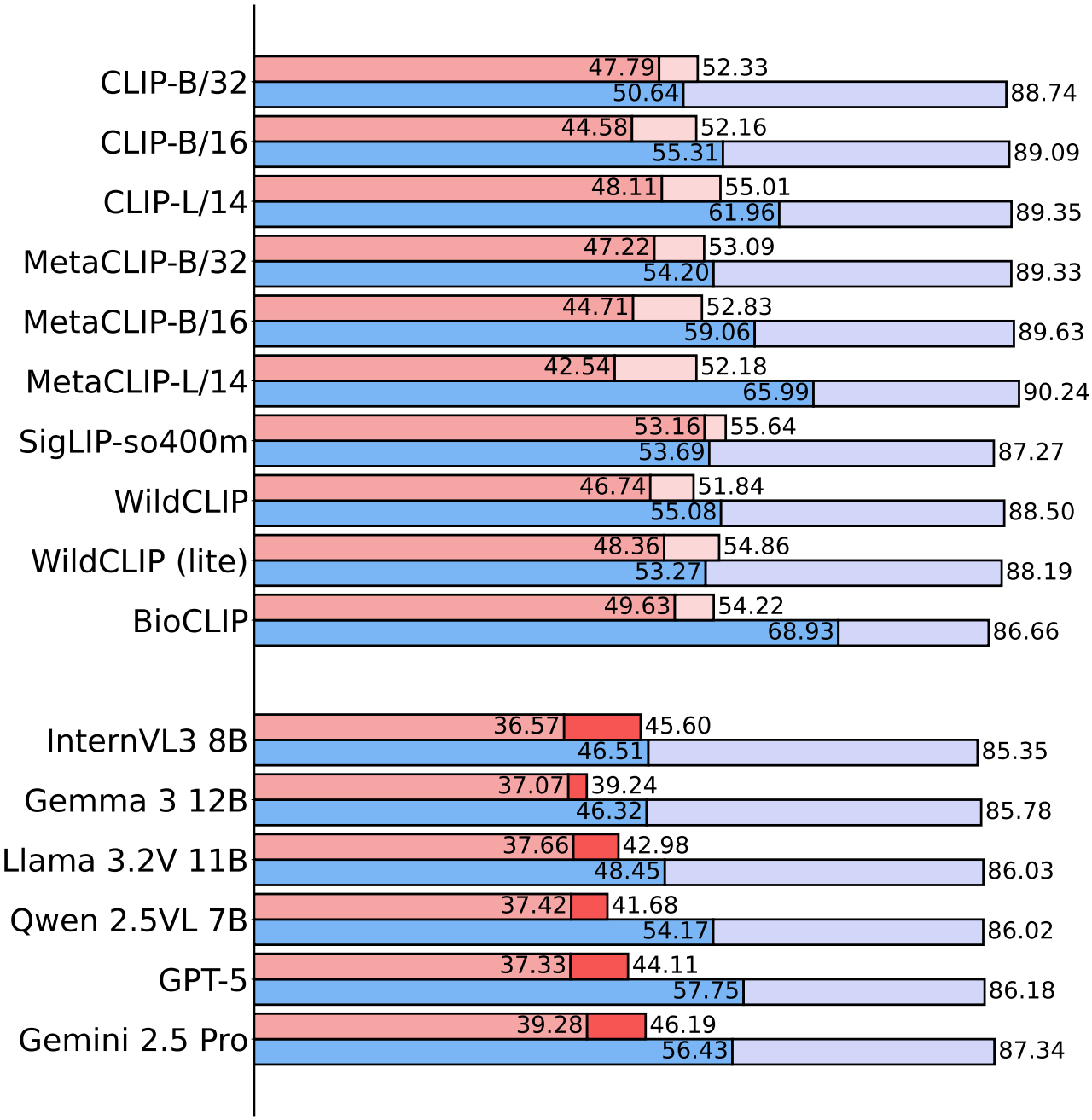


Figure 11. **Summary of classification and abstention performance on RealBirdID.** Blue bars show classification performance (IG), a metric which mixes species and genus-level accuracy over coverage on answerable (A) examples, while red bars show abstention capability, measured as AUROC for separating answerable (A) from unanswerable (UA) instances (higher is better for both). The lighter bars correspond to increases in performance from using *species range maps* whereas the darker bars indicate performance decreases. Notably, abstention tradeoff performance decreases for MLLMs.

Figure 12. **100 examples of images from the answerable set.** The answerable set is sampled from Research Grade iNaturalist images to fill out species corresponding to sampled Unanswerable data.

Figure 13. **30 images with *angle/occlusion* abstention reasons from the unanswerable set.**

Figure 14. **30 images with *vocalization* abstention reasons from the unanswerable set.**

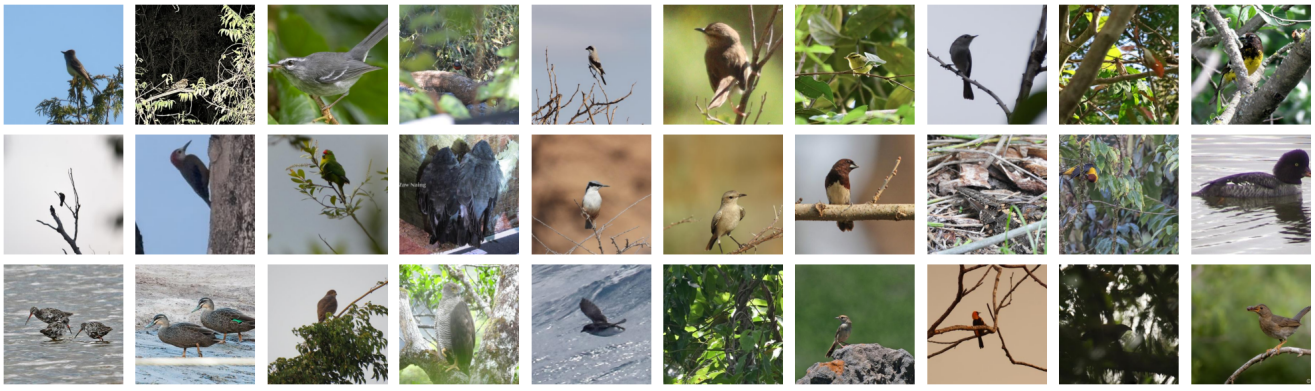


Figure 15. **30 images with *image quality* abstention reasons from the unanswerable set.**