CAN WEIGHT-SPACE PROGRAMMING SURFACE THE DOMAIN EXPERTISE HIDDEN IN A SMALL VISION-LANGUAGE MODEL? EVIDENCE FROM CAMERA TRAPS AND GALÁPAGOS SPECIES

A PREPRINT

Andrew Hill

ABSTRACT

AI editors rewrote the weights of a small vision-language model, LLaVA-OneVision Qwen2-0.5B, directly, without gradients, optimiser updates or adapters. The study has three parts, and the third compares that direct route with a LoRA fine-tune. In the first part, a general camera-trap edit raised balanced accuracy from 64.09% to 81.76% on 1,611 crops from 45 cameras, with each camera's label-derived choices (gain 17.67 points; 95% interval 12.41 to 23.61). On a new camera-trap collection with an 11-name list, restricted to in-list animals, it closed about 41% of the gap to SpeciesNet, a purpose-built camera-trap classifier (27.92% to 55.55%, against 95.13%). Open naming made little progress, and the small model collapsed under a 64-name list before and after editing. A separate SigLIP readout of the model's own image features scored at or above the edited answers throughout this part. On the same 1,611 crops the edited model's 81.76% is level within noise with BioCLIP 2, an open image model built for species recognition across the tree of life, which scored 84.07% without adaptation ($+2.31$ points, 95% interval -1.58 to $+6.38$). The second part aims the same construction at animals SpeciesNet cannot name. On Galápagos citizen-science photos, the SigLIP readout separates native iguanas from green iguanas, while the model's answers call almost every iguana native and SpeciesNet has no label for either native iguana genus. An island edit was chosen on a separate selection set and scored once, with its weights fixed, on a held-out test set of 697 crops from locations used in neither fitting nor selection. Mockingbird-versus-brown-noddy balanced accuracy rose from 70.83% to 98.33%, equal to the SigLIP readout and against 3.33% for SpeciesNet. Land and marine iguanas against green iguanas rose from 50.00% to 81.22% and 80.37%, above the readout's 78.29% and 77.45% and against 12.44% for SpeciesNet. Each edit is a specialist: on the camera benchmark the island edit scored 72.14%, 8.05 points above the unedited model and 9.62 below the general edit. The third part fits one model for both jobs, starting from the general edit and using the same 2,902 practice crops. A LoRA fine-tune reached 87.58% and 91.69% on the land- and marine-iguana contrasts, 93.33% on the birds and 88.20% recall of invasive rats, while scoring 79.22% on the camera benchmark. A direct combined edit, fitted in closed form, scored 80.70% on the camera benchmark but only 55.49%, 55.49%, 90.83% and 17.70% on the same island tasks. The fine-tune beat the combined edit on every island task and matched it on the camera benchmark within noise (-1.48 points, 95% interval -5.80 to $+2.54$). The direct edit was the faster route: about an hour to build, against about three hours for the fine-tune.

1 Introduction

The study asks whether a stronger AI can directly edit an off-the-shelf model's weights so that its answers improve, and whether that edit can be aimed at animals a purpose-built classifier cannot name. It uses one small vision-language model and one family of edits, and has three parts. The third asks whether one model can hold both jobs, and compares the direct edit with a standard fine-tune.

The first part is a general camera-trap edit. AI editors raised the model's balanced accuracy from 64.09% to 81.76% on the original 45-camera benchmark, where each prompt listed that camera's choices. On a new collection with a fixed 11-name list, the same edit raised the model from 27.92% to 55.55% against 95.13% for SpeciesNet, a purpose-built

camera-trap classifier. The gap fell from 67.21 to 39.58 percentage points, about 41% closed. That comparison scores only crops whose true species is on the list; 501 valid out-of-list crops were not evaluated under the 11-name condition. Without a list the edited model scored 4.74% against SpeciesNet's 77.29%.

The second part starts from a measurement. A separate SigLIP readout of the model's own image features scored at or above its edited answers throughout the first part. We therefore asked where that readout could see something that neither the model's answers nor SpeciesNet expressed. On Galápagos citizen-science photos, the readout separated native land and marine iguanas from the introduced green iguana, and Galápagos mockingbirds from other birds. The model's answers called almost every iguana native. SpeciesNet has no label for either native iguana genus and none for the Galápagos mockingbird species.

The second part is a targeted island edit, built with the same closed-form construction and scored on a held-out, location-split test set. Mockingbird-versus-brown-noddy balanced accuracy rose from 70.83% to 98.33%, equal to the readout and against 3.33% for SpeciesNet. Land and marine iguanas against green iguanas rose from 50.00% to 81.22% and 80.37%, above the readout's 78.29% and 77.45% and against 12.44% for SpeciesNet. The island edit gave up part of the first edit's camera-trap gain.

The third part fits one model for both jobs. Two models start from the general edit and learn from the same 2,902 practice crops, camera and island together. A LoRA fine-tune, a standard gradient method, reached 87.58% and 91.69% on the land- and marine-iguana contrasts, 93.33% on mockingbirds against brown noddies and 88.20% recall of invasive rats, while scoring 79.22% on the camera benchmark. A direct combined edit, fitted in closed form like the first two, scored 80.70% on the camera benchmark and 55.49%, 55.49%, 90.83% and 17.70% on those island tasks. The fine-tune beat the combined edit on every island task and matched it on the camera benchmark within noise. The direct edit took about an hour to build and the fine-tune about three.

The contributions are:

- a gradient-free edit of the projector and output rows that repaired camera-trap answers under short supplied lists, with a component analysis in which the answer-format repair was the largest single lever;
- a transfer test on a new collection against SpeciesNet, which locates the repair's useful scope in short lists and shows where it stops helping;
- a measurement showing that the model's visual features separate island natives that its answers and a purpose-built classifier do not;
- a targeted edit for those natives, chosen on a separate selection set and scored once on held-out photos, which matched or exceeded the visual readout on every native contrast tested, with its cost measured on the original camera benchmark;
- a comparison of one model for both jobs, in which a LoRA fine-tune beat a direct combined edit on every island task and matched its camera score, while the direct edit was about three times faster to build.

2 Context and related work

The approach follows weight-space programming: a general-purpose AI agent inspects neural weights, uses numerical tools and forward diagnostics, and writes changes directly into those weights [1]. We applied that idea to an animal-naming task. Numerical fitting from practice examples was allowed; backward passes, autograd and optimiser updates were disabled. The resulting inference still used the native edited model to generate its answer.

LLaVA-OneVision connects an image encoder to a language model with a two-layer multilayer perceptron, or projector [2]. The model studied here combines a SigLIP vision tower [3] with Qwen2 language weights [4]. Comparable encoder, connector and language-model components appear in other open families, including Qwen2-VL and InstructBLIP [5, 6]. This common structure makes the connector a place to test the repair question in other models.

SpeciesNet provides a different applied comparison: a purpose-built camera-trap classifier, distributed with native species and geographic reasoning components [7, 8]. The original images came from Caltech Camera Traps and Island Conservation Camera Traps [9, 10]. The new collection came from Felidae Conservation Fund [11, 12]. We use the supplied labels and frozen mappings described below; the comparison scores each system on the same selected crops.

3 Study design

All three parts use the model, prompt form and scoring form described here. The direct edits use the construction in Section 4. The camera-trap data and the successive editing process in this section belong to the first part. Section 6 gives the second part's photos, split and selection, and Section 7 gives the third part's two models.

3.1 Model, inputs and reference labels

The model was the public `lmms-lab/llava-onevision-qwen2-0.5b-ov` checkpoint, converted to the standard Transformers `LlavaOnevision` representation in float32. Its architecture, tokenizer, prompt, decoding and answer scorer stayed fixed during editing. All native weight tensors were readable and editable. The final delivered state changed three projector tensors and one tied language input/output table. The native vision blocks and language attention/MLP weights were unchanged in that state.

The original evaluation contained 1,611 fixed animal crops from 45 cameras: 27 Caltech and 18 Island Conservation cameras. Each image prompt supplied that camera's own list of animal names. Those lists were derived from labels at the camera, so they provided information about which animals occurred there. The prompt asked for the animal's name from the listed choices and no other text. The frozen parser accepted a unique listed name, an option letter, or an agreeing letter/name pair; it rejected conflicting pairs and extra prose. An unaccepted response counted as incorrect. The benchmark therefore measured a complete decision under supplied choices, including whether the answer was usable. Crops used the frozen bounding box with 25% padding per side, clipped to the frame. Generation was greedy, with one beam, at most 1,024 new tokens and a repetition penalty of 1.1; these settings were unchanged between the original and edited model.

The editor could use 23,139 labelled practice crops from 57 cameras. Study software reserved 204 separate private tuning crops. Whole-image, burst and exact-hash groupings did not cross the practice, tuning and session-test boundaries. The test cameras were withheld from this session's practice, but came from earlier screening. The practice vocabulary contained 58 candidate names; not every name had labelled practice instances. The original scored prompts had 2–12 options.

The separate visual reference ran the native vision features through the original SigLIP final block, normalisation, pooling and text readout, using the global image view. It selected a label from the same camera list. This is an additional inference route. The editor could inspect the reference components and use reference outputs on disclosed practice examples.

3.2 Successive editors and checkpoint selection

Four successive editing sessions used Claude Fable 5.2, Claude Opus 5.5, GPT-6 Astra Pro, then Claude Fable 5.2. Each inherited the accumulated weights, work products and notes. These are successive parts of one editing process, not independent trials. The first three editing sessions contributed the checkpoints scored at 77.56%, 76.48% and 77.14% on the original session-test cameras.

A persistent numerical workspace exposed the model's stored arrays. Editors could analyse disclosed examples, inspect activations and token scores through forward tools, solve numerical systems and replace weight values. Previous compatible patches were available as optional starting points. The editors completed 39 numbered checks. Each check evaluated a saved weight state by generating answers on practice crops; its number identifies that candidate evaluation. The tools permitted numerical fitting from labels and the visual reference; the method was gradient-free, not label-free or fitting-free.

Study software evaluated candidate checkpoints on the 204 private tuning crops. It updated the selected state only when the new observed score was higher; ties retained the incumbent. Tuning labels, outputs, scores and ranks were not returned to the editor's workspace. Check 39 scored 86.5728% on this selector and check 38 scored 85.8784%. The frozen analysis retained distinct states within one percentage point of the selector's best. The selector supplies an observed ordering; the 204-crop reservation leaves the quality ordering of these near-tied states uncertain, so both are reported.

Test scores were inspected at several stages and informed decisions to continue editing; check 39 remained the private-tuning-selected delivered state.

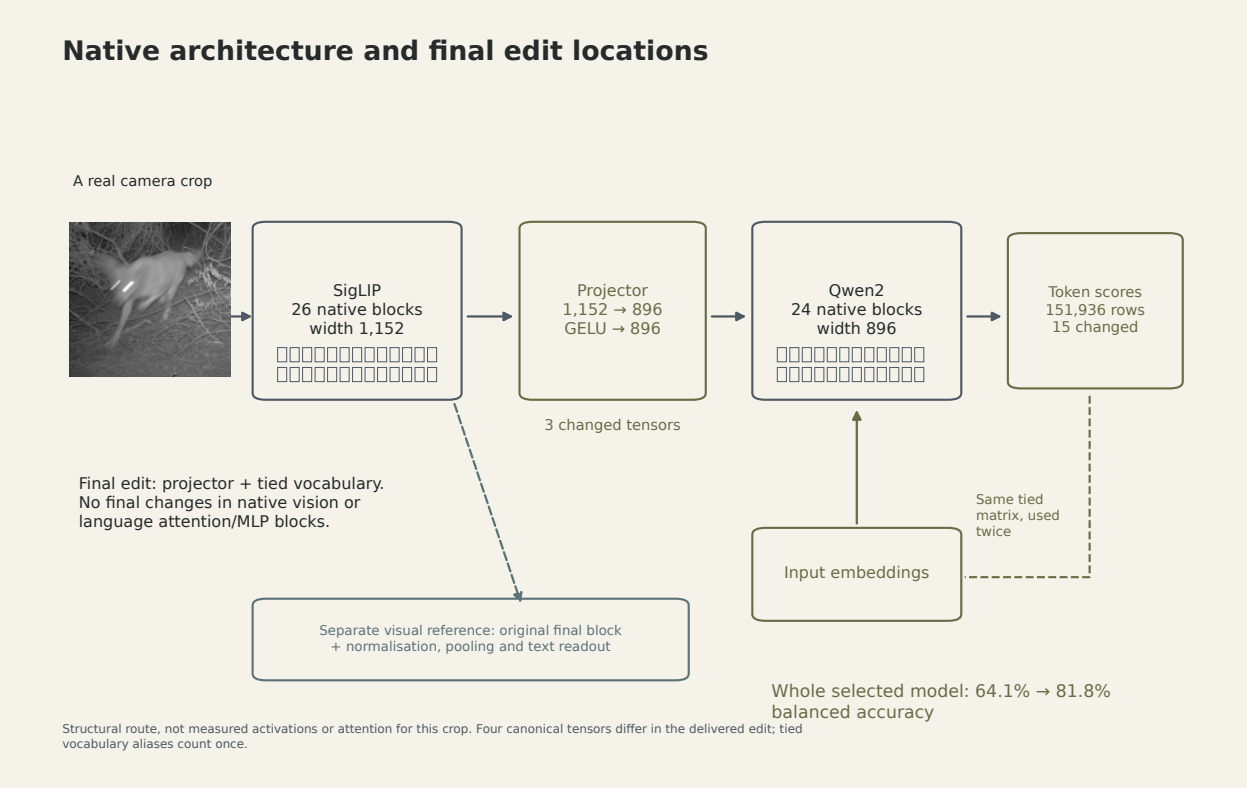


Figure 1: The native image-to-answer path and delivered edit locations. The final state changes the projector and tied input/output table; the vision and language attention/MLP blocks are unchanged. Structural arrows show the architecture, not a measured crop-specific attention map. The separate SigLIP branch adds its original completion and readout.

3.3 Scoring and uncertainty

For a source camera or location g, let $\mathcal{C}_g$ be the reference labels present and n_{gc} the number of scorable crops with label c. With y_i the reference and $\hat{y}_i$ the accepted answer, the primary score was

$$\mathrm{BA} = \frac{1}{G}\sum_{g=1}^{G}\frac{1}{|\mathcal{C}_g|}\sum_{c\in\mathcal{C}_g}\frac{1}{n_{gc}}\sum_{i:y_i=c,\, i\in g}\mathbf{1}[\hat{y}_i = y_i]. \tag{1}$$

Thus each source unit counts equally, and labels present within that unit count equally. Pooled plain accuracy weights crops equally and is reported separately. Fixes and harms count individual crops, not scoring weight.

Intervals for paired changes used 2,000 shared resamples of whole source units, preserving the paired predictions. The original study resampled cameras; the Felidae comparison resampled source locations. These intervals describe the sampled mixture for fixed states. They do not estimate variation across independent editor sessions, remove adaptive test reuse, or provide simultaneous multiplicity-adjusted claims for every displayed contrast.

4 Editing the path from image to answer

This section describes the general camera-trap edit, check 39. The island edit in Section 6 reuses its hint mechanism and letter-row correction with new class lists. The combined edit in Section 7 keeps check 39's hints and answer rows, and adds island hints in projector units that check 39 had not changed.

4.1 Finishing a concise answer

A correct animal description was sometimes unusable as a decision. A saved fox example began "B. The image shows a fox" and continued into a paragraph, while a bare "B" would have been accepted. The editor diagnosed competition between answer tokens and common introductory words. The editor attributed that competition partly to the fixed

repetition penalty: it disadvantaged answer letters and the end-of-turn token present in the prompt, relative to the continuation tokens `The` and `space+The`.

In the delivered model, rows for `The` and `space+The` became zero, and the end-of-answer row was multiplied by 1.5. A token's direct output score is the dot product of its row with the final language vector. A zero row gives a zero direct score, rather than zero probability. Scaling the end row increases a positive dot product and can help generation end after a letter; it is not a universal probability increase. Because the table is tied to input embeddings, changing these rows also changes their input representations where used.

The full edit reduced unusable original-test replies from 237 to zero. The restoration analysis provides a narrower measurement: restoring the three format rows in the full edited model brought back 212 unusable replies and lowered balanced accuracy to 72.03%.

4.2 An image-dependent hint in existing projector units

For an image-feature vector z, the native projector computes

$$p(z) = W_2\ \mathrm{GELU}(W_1 z + b_1) + b_2. \tag{2}$$

The editor fitted visual maps from disclosed practice features, labels and reference representations. It placed class-score coefficients, biases and position-sensitive terms into existing rows of W_1. The corresponding outgoing columns in W_2 supplied animal-name embedding directions. A unit's response therefore controlled how strongly a name-like vector was added to the image-token sequence. The added vector entered the native image-token sequence.

Two existing base-view token addresses, 568 and 569, carried the preceding and final token parts of an animal name. These are addresses in the 27-by-27 image-token sequence, not locations of the animal in the photograph. The final circuit reassigned 120 existing hidden units: 58 class-name units at each address and two pairs of cancellation units. The 120 rewritten hint units are included in the 547 changed incoming rows. Other changed rows received a position-sensitive subtraction to suppress their existing responses at the final hint address. The recipe excludes the rewritten hint units from that suppression. Cancellation removed a constant output offset at one address and an unwanted residual vector at the other. Saved unit and row maps show the physical changes in Figures 1–2.

A deliberately false rabbit hint provided a practice diagnostic of steering. Eighteen of 32 non-rabbit practice crops produced rabbit after the intervention; none of the matched original replies did. This showed that the channel could influence the answer.

4.3 From an animal name to an option letter

A useful animal-name direction still had to select the correct letter in the current list. The final correction fitted residual directions from labelled practice language states and added four times those directions to the existing A–K rows. In representative notation, the residual fit solves a ridge system of the form

$$r_j = (H^\top D_j H + \lambda_j I)^{-1} H^\top D_j (y_j - p_j), \qquad w'_j = w_j^{\mathrm{intermediate}} + 4 r_j, \tag{3}$$

where H contains practice states captured from preceding candidates, D_j selects examples offering letter j, y_j is its correct-answer indicator and p_j its within-option probability under the intermediate head. The final call uses $\lambda_j = 0.1\,\mathrm{tr}(H^\top D_j H)/896$, excludes single-option examples and skips a letter with fewer than five eligible states. The correction is added to the intermediate head, rather than directly to the public base rows. The packaged recipe specifies the state banks and numerical precision.

Only 15 rows of the vocabulary table changed: A–K, `space+A`, two "The" rows and the end-of-answer row. A–K change norms ranged from about 0.09 to 0.15 beside original row norms around 0.45; `space+A` was heavily rewritten for the single-option case. No letter row after K changed. Figure 3 plots the saved row norms. These output edits also alter input embeddings, so a restoration is not a pure intervention on the final logits with an otherwise fixed state.

5 Part one: a general camera-trap edit

5.1 Original camera benchmark and adjacent checkpoints

Table 1 reports the original benchmark. Check 39 improved balanced accuracy by 17.67 points and corrected 349 crops while making 33 previously correct crops wrong. Of the original 237 unusable replies, 209 became correct and 28 became accepted wrong choices; a further 140 accepted wrong animal decisions became correct. The delivered result therefore includes both response-format repair and changed decisions.

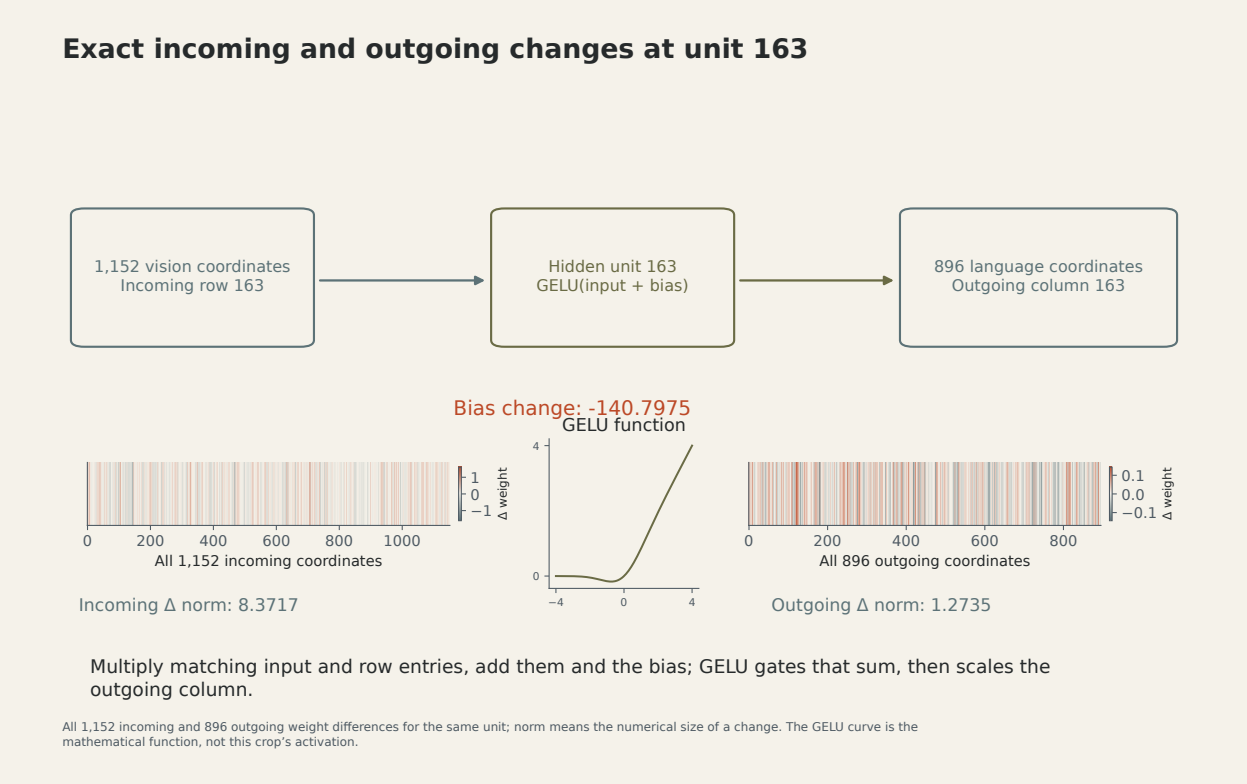


Figure 2: Saved numerical differences at one existing projector unit, 163, with its incoming row, bias and outgoing column. Its animal-specific semantic role is not established by the excerpt. The GELU curve is the mathematical function, not a measured activation for the displayed crop.

Check 38 scored 81.32%, with a gain of 17.23 points and a paired-camera interval of 11.93 to 23.12. It remains a sensitivity row beside check 39, not a replacement selected using the test result. The first three editors' checkpoints cluster around 77%, but inheritance and shared adaptive decisions prevent treating them as independent demonstrations of convergence or estimates of editor ability.

State	Balanced (%)	Plain (%)	Gain, points [95% interval]
Unedited	64.09	64.25	N/A
Editor stage 1 / check 8	77.56	80.38	+13.47 [8.36, 19.17]
Editor stage 2 / check 12	76.48	77.59	+12.39 [6.85, 18.53]
Editor stage 3 / check 25	77.14	80.94	+13.05 [8.17, 18.33]
Editor stage 4 / check 38	81.32	83.36	+17.23 [11.93, 23.12]
Editor stage 4 / check 39	81.76	83.86	+17.67 [12.41, 23.61]
SigLIP reference	82.79	86.28	N/A

Table 1: Original with-list benchmark, 1,611 crops from 45 cameras. Gains are paired against the same unedited predictions; intervals resample whole cameras. These cameras were held out from this session's practice but previously screened and scored at several development stages. SigLIP uses an additional visual readout.

5.2 The largest individual restoration effect is format

Restoring the original output-format rows lowered the full score from 81.76% to 72.03%, a change of -9.73 points with a paired-camera 95% interval of -15.22 to -5.06. This was the largest drop among the individual parameter groups tested. It also restored 212 non-answers. The least elaborate part of the repair had a large effect on the delivered task.

The two broad halves were close to additive at this level: output-row edits alone gained 11.07 points and the hint channel alone 6.07, summing to 17.14 beside the combined 17.67. The residual is 0.53 points, with no interval estimated for it. Their absolute scores were 75.16%, 70.16% and 81.76%. Restoring the previous-token branch lowered the full score by

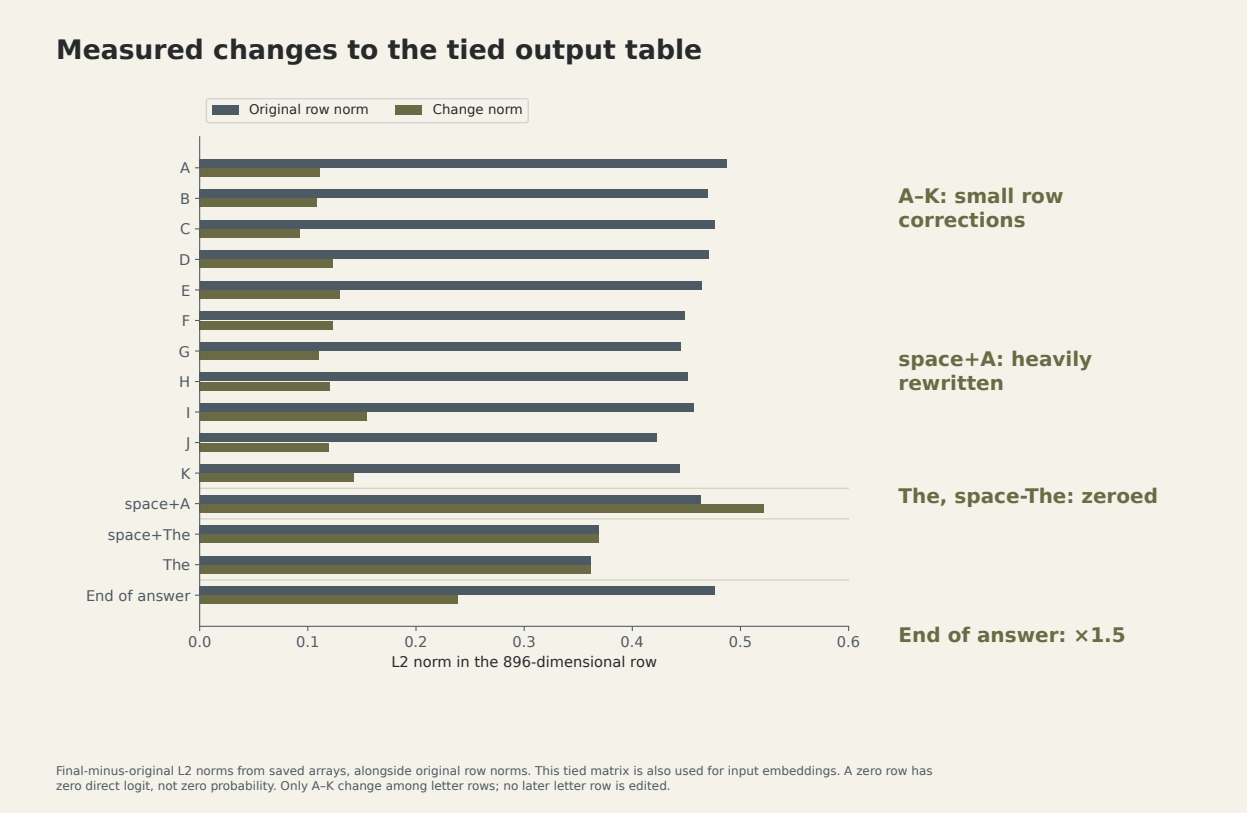


Figure 3: Original row norms and final-minus-original change norms for all 15 changed rows. Only A–K are corrected among option letters. Two "The" rows are zeroed, the end-of-answer row is scaled by 1.5, and the space-A alias is heavily rewritten.

6.49 points and restoring the last-token branch by 5.48. Restoring the letter rows lowered it by 1.37 points, with an interval including zero. Restoring the single-option alias left correctness and accuracy unchanged on this benchmark, which had no single-option prompt.

These are conditional original-value restorations, not additive shares of a gain. The output table is tied to input embeddings. The previous-token and last-token branch restorations retain that suppression of other rows at the final hint address. The controls restore parameter groups rather than isolated semantic functions. Table 2 and Figure 4 retain every state.

State	Balanced (%)	Non-answers	Change from full, points [95% interval]
Full edit	81.76	0	+0.00 [+0.00, +0.00]
Restore single-option alias	81.76	0	+0.00 [+0.00, +0.00]
Restore letter rows	80.39	0	-1.37 [-3.54, +0.68]
Restore previous-token branch	75.27	0	-6.49 [-8.87, -4.45]
Restore last-token branch	76.29	0	-5.48 [-7.84, -3.33]
Restore format rows	72.03	212	-9.73 [-15.22, -5.06]
Output-row edits only	75.16	0	-6.60 [-9.00, -4.56]
Hint channel only	70.16	217	-11.60 [-17.25, -6.58]

Table 2: Conditional restorations on the original with-list cohort. "Restore" returns that group to its original values while retaining the rest of the delivered edit. Non-answers are generated responses rejected by the frozen scorer, not inference-process failures.

5.3 Improvement was smaller without supplied choices

Removing the camera option list reduced both original and edited scores. Under strict no-list scoring, check 39 improved balanced accuracy from 34.88% to 41.19%, a gain of 6.31 points with a paired-camera interval of 3.23 to 10.72. Taxonomy-lenient scores changed from 47.78% to 49.73%, a gain of 1.94 points with an interval of 0.62 to 3.45.



Figure 4: All conditional component restorations and the two component-only states. Restoring the format rows gives the largest individual drop tested. Restorations of tied input/output weights and shared suppression are not additive shares of the gain.

The original no-list lenient rule allowed declared ancestor or descendant relationships; an accepted descendant of a coarse reference did not establish that its finer species identification was correct.

All five edited frozen states improved aggregate no-list scores over the unedited baseline (Table 3). Check 12 had the highest strict no-list score among these states, but no checkpoint was reselected using this comparison. The with-list and no-list rows differ in supplied information and answer form. Their gains do not divide into a unique list-dependent and visual contribution.

State	With list	No list, strict	No list, lenient
Unedited	64.09	34.88	47.78
Check 8	77.56	39.19	49.24
Check 12	76.48	45.51	52.47
Check 25	77.14	39.47	49.23
Check 38	81.32	41.32	49.41
Check 39	81.76	41.19	49.73

Table 3: Original benchmark, all six fixed states. Entries are equal-camera balanced accuracy (%) on the same 1,611 crops / 45 cameras. Strict and taxonomy-lenient mappings were frozen; unresolved output strings receive no credit.

5.4 A new collection: repair transfers under the short list

The frozen comparison on Felidae Conservation Fund images retained five arms: unedited, checks 39 and 38, the separate SigLIP reference, and SpeciesNet. Lists were fixed from dataset-wide metadata and class frequencies in metadata recorded before the capture cut-off, before any model outputs. The 11-name list was not constructed from the labels of these scored crops. The full 64-name list was the predeclared primary condition. Eligible captures were later than the package-release cut-off stated in the abstract and temporal appendix, while the model training cut-off remains unpublished.

The frozen sample selected single-category labelled animal images with known source locations. It kept at most one image per explicit sequence or, where absent, per same-location 60-second burst, with caps of 80 images per location, 30 per location/class and 600 per class. It retained the highest-confidence MegaDetector v5a.0.1 animal detection at confidence at least 0.1 and cropped that box without padding. All five arms received the same pixels. Table 6 separates detector-valid crops, scorable crops and the short-list restriction. Source location IDs were grouped where an explicit shared physical station was known; their remaining physical independence is unverified. The score describes classification conditional on these selected detections and supplied provider labels. It does not estimate end-to-end detection accuracy or natural-prevalence performance.

The narrow-list scoring subset contained 1,855 crops at 74 source locations, all with a reference animal label in the 11-name list; 501 valid out-of-list crops were not evaluated under the 11-name condition. Check 39 improved balanced accuracy from 27.92% to 55.55%, a gain of 27.64 points with a paired whole-location interval of 21.16 to 34.35; there were 503 fixes and 30 harms. Check 38 scored 55.48%. The matched SigLIP reference scored 56.41%, while SpeciesNet scored 95.13%. The delivered-task repair thus transferred under the tested short list and again approached the visual reference. The gap to SpeciesNet fell from 67.21 to 39.58 points, about 41% closed (Table 5).

The full 64-name primary condition exposed a limit already present in the small base model. It contained 2,355 scorable crops at 75 locations. The unedited model scored 2.22%, check 39 scored 0.27% and check 38 scored 0.35%. SigLIP scored 43.79% and SpeciesNet 90.55%. Table 4 and Figure 5 report both.

The same 1,855 narrow-list crops also had full-list results. Check 39 scored 0.47% with the full list and 55.55% with the narrow list. This matched comparison removes the crop-cohort difference. Length, label contents and serialised option positions still changed together, so it does not identify a cardinality-only effect or a universal safe list length.

Arm	Full 64, all	Full 64, 11-name cohort	Narrow 11, same cohort
Unedited	2.22	3.25	27.92
Check 39	0.27	0.47	55.55
Check 38	0.35	0.61	55.48
SigLIP reference	43.79	40.90	56.41
SpeciesNet	90.55	93.74	95.13

Table 4: Felidae collection: equal-location balanced accuracy (%). 64-name list, all crops uses 2,355 scorable crops at 75 locations; both matched-cohort columns use the same 1,855 scorable crops at 74 locations. Every arm shares the relevant mask. 64-name list, all crops remains primary.

Condition	Gap before	Gap after	Change in VLM
11-name	67.21	39.58	+27.64
64-name	88.33	90.28	-1.95
No-list strict	73.27	72.55	+0.72
No-list lenient	70.19	70.80	-0.61

Table 5: Distance to SpeciesNet on the new collection, percentage points. Gaps subtract the displayed two-decimal scores; the fraction closed uses unrounded scores (41.12% under the 11-name list). The strict no-list gain closes less than 1%; lenient naming and the full list widen the gap.

Stage	Full list	11-name list
Sampled images	2,426	*subset of full sample*
Detector-valid crops (answer forms)	2,402	1,901
Unresolved animal labels	47	46
Scorable crops (accuracy)	2,355	1,855
Source locations in accuracy score	75	74

Table 6: Felidae cohort flow. Twenty-four sampled images lacked a usable detection. The 501 detector-valid crops outside the 11-name list were evaluated under the full-list condition and excluded from the 11-name condition.

5.5 Open naming and answer forms on the new collection

The Felidae no-list comparison was less favourable to the VLMs (Table 7). Check 39 scored 4.74% strict and 19.34% lenient, versus 4.02% and 19.95% unedited. In this comparison, strict requires taxonomic correctness at the dataset

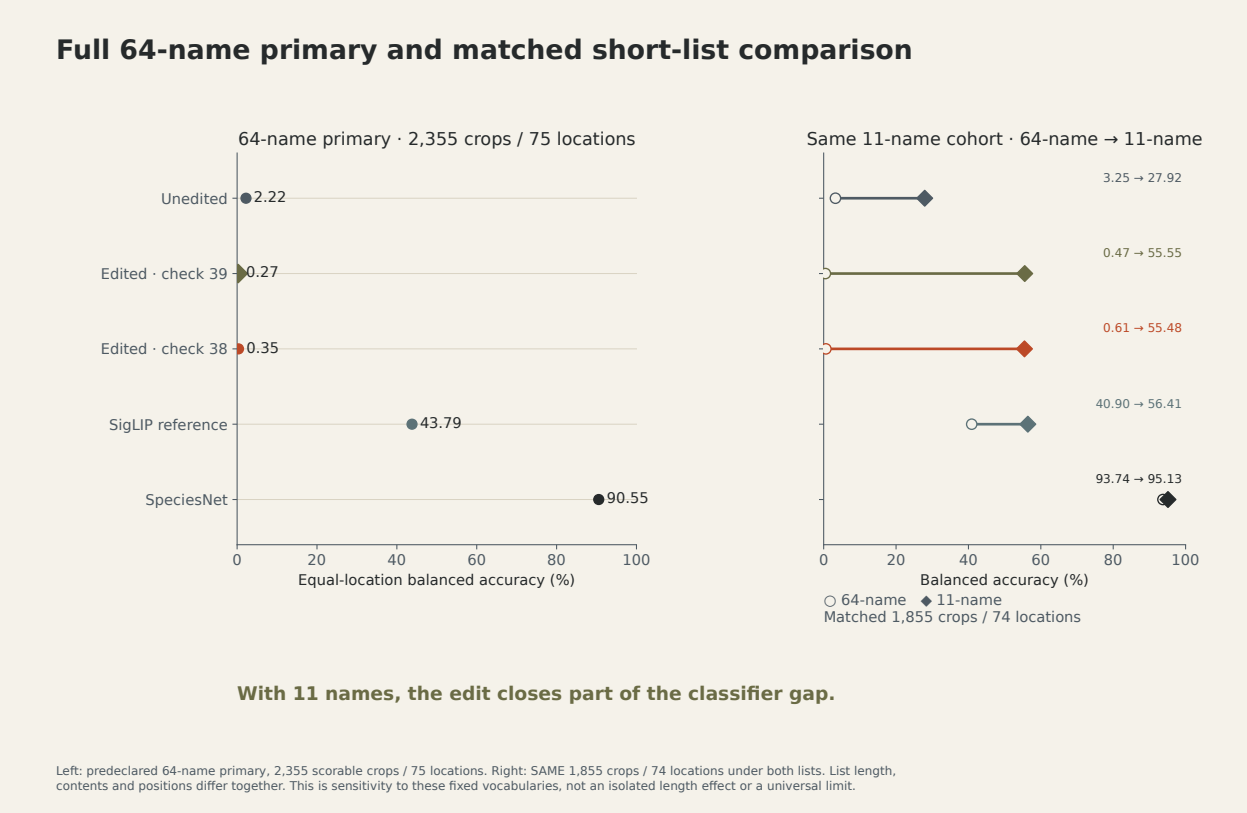


Figure 5: The primary 64-name condition on all scorable Felidae crops, and both lists on the same 11-name cohort. The short-list repair transfers, while SpeciesNet remains much more accurate. The matched comparison removes cohort differences but still changes list length, contents and option positions together.

label's specificity; the fixed lenient rule also accepts an appropriate higher genus, family or order. SpeciesNet used native USA/CA geographical priors, which the VLM prompts did not supply. SigLIP searched a frozen finite dictionary rather than generating unrestricted text. Those are method differences, not interchangeable open-naming systems.

Arm	No list, strict	No list, lenient
Unedited	4.02	19.95
Check 39	4.74	19.34
Check 38	4.73	20.43
SigLIP reference	13.60	32.25
SpeciesNet	77.29	90.14

Table 7: Felidae no-list results on 2,355 scorable crops / 75 locations, equal-location balanced accuracy (%). SpeciesNet includes native geographical priors; SigLIP is a finite-dictionary reference.

Raw answer forms help interpret the gain. Under the short list, check 39 produced 1,900 bare letters and one agreeing letter/name combination, while the unedited model produced 1,541 agreeing and 360 conflicting combinations. These form counts cover all 1,901 valid crops, including unresolved labels; the accuracy denominator is smaller. Under the full list, check 39 produced 2,152 bare letters but scored only 0.27% balanced accuracy. A concise answer is not necessarily a correct animal decision. Conflicting names and letters were rejected.

The gain varied across labels. Gray fox declined from 92.76% to 82.48%, while mule deer improved from 0.16% to 67.58%. The mouse-or-rat option had no scorable examples and remains N/A. Position and species identity are confounded, and overlap with practice labels depends on the recorded taxonomic mapping. Tables 13 and 14 give all answer forms and short-list labels.

6 Part two: island animals a camera-trap classifier cannot name

6.1 The model's features separate natives its answers do not

The second part starts from a measurement on Galápagos animals. The unedited model, check 39, the SigLIP readout and SpeciesNet scored the same research-grade iNaturalist photos of native animals and their look-alikes [13]. The three list-based arms chose from a fixed 11-name list for each animal group, drawn from regional checklists [14]; SpeciesNet kept its full vocabulary. Native-versus-look-alike balanced accuracy gives locations equal weight within each direction, then gives the native and look-alike directions equal weight.

On native land or marine iguanas against the same 88 green iguanas, the readout scored 79.94%. The unedited model and check 39 scored 50.00% and 52.50%: they called almost every iguana native, recognising 0.00% and 5.00% of green iguanas. On Galápagos mockingbirds against other birds, the readout scored 97.66%, check 39 75.55% and the unedited model 63.21%. The readout's exact-genus recall was 95.00% for land iguanas, 96.48% for marine iguanas and 95.68% for mockingbirds. SpeciesNet's standard output scored 12.37% on the iguana contrasts and 0.88% on the birds. It has no label for either native iguana genus. It has the genus *Mimus* through two mainland species but none of the four Galápagos species, and its outputs on these photos were mostly "reptile", "bird" or "animal". Native rice rats had only four photos, which every list-based arm assigned to the wrong native genus.

For these animals the information is present in the model's image features, its answers do not use it, and the purpose-built classifier has no label to give. The second part tests whether the construction of Section 4 can carry that information into the answers.

6.2 Plan, photos and the location split

Before any edit or scoring, a target paper set out the abstract with blank numbers, the abstract that would be reported if the edit failed, and the assumptions most likely to break it. A non-author reviewer then checked the target paper and design. Its findings, including a correction that narrowed the location-separation claim, were resolved before editing began. No numerical success bar was set.

The island edit uses the measurement's sources, crop rules and scoring rules, with newly sampled images. Whole public location groups were assigned to practice, selection or test before any model output. Galápagos photos used county identities across taxa: Isabela for practice, San Cristóbal for selection and Santa Cruz for test. Other photos used unobscured 0.1-degree cells, and camera-trap controls used camera identities. Selection and test excluded every image and photo event sampled for the measurement above, and observations of the same species by the same observer on the same day. All 45 original benchmark cameras were excluded from every role. After download failures and crops without an accepted animal detection, 846 practice, 426 selection and 697 test crops remained (Table 8).

Three fixed 11-name menus, for rodents, iguanas and birds, list native taxa, look-alikes and controls in a fixed order. In the test set, the native directions held 60 land iguanas, 59 marine iguanas and 60 mockingbirds, all from Santa Cruz. The green-iguana direction held 57 crops from 41 locations and serves both iguana contrasts. Four other-bird species were eligible for the bird contrast, but only brown noddies (*Anous stolidus*, 60 crops) reached the test. No fresh native rice-rat sample was available.

Role	Sampled	Benchmark-camera exclusions	Eligible	Fetch failures	No animal detection	Crops
Practice	930	35	895	22	27	846
Selection	461	5	456	15	15	426
Test	780	28	752	19	36	697

Table 8: Island study cohort flow. Location groups were assigned to roles before any model output. Crops are the shared animal crops every arm received.

6.3 Building and choosing the island edit

The Codex research assistant (OpenAI GPT) wrote programmes that fitted the edit in closed form, as for check 39: ridge regressions from labelled practice features and SigLIP reference scores, written into projector weights and answer-letter rows. There were no gradient updates, no optimiser and no LoRA adapters. Two starting versions both reused check 39's two-position name-hint mechanism. One kept check 39's language and output weights and its hints for the original 58 camera classes. The other used the original language and output weights and removed those hints. Both added hints for the island menus. For each version, a fixed schedule tried two ways of encoding a name hint and three hint

magnitudes. The best candidate from each version then received two strengths of letter-row correction, giving sixteen edited candidates.

The 426 selection crops chose among them. The selection score averaged native-status balanced accuracy and a stricter version requiring the true native genus, over the two contrasts complete in the selection set: marine against green iguanas, and mockingbirds against other birds. The chosen candidate started from the original language and output weights, encoded each hint as a short distinguishing name such as "land iguana", used hint magnitude 0.6, and applied four times the fitted letter correction. Its weights were fixed before any test or camera-benchmark output was produced. Logs of the editor's file reads, 98,499 recorded open events, show no read of the test records.

6.4 Results on the held-out test

Table 9 reports the test set. Either native iguana genus counts as a native answer; a green-iguana answer must name the genus *Iguana*. A mockingbird answer must name *Mimus*, and a brown-noddy answer must resolve to a bird outside *Mimus*.

The island edit raised native-versus-look-alike balanced accuracy by 31.22 points for land iguanas, 30.37 for marine iguanas and 27.50 for mockingbirds against brown noddies. The iguana gains came from the look-alike side. The unedited model answered native for every green iguana, 0 of 57 correct. The island edit got 36 of 57 right while keeping 60 of 60 land iguanas and 58 of 59 marine iguanas native. On the bird contrast it named 60 of 60 brown noddies correctly, against 26 of 60 unedited, and kept 58 of 60 mockingbirds. Check 39, fitted for camera traps, scored 51.83% on each iguana contrast and 91.67% on the birds.

The island edit reached or exceeded the SigLIP readout on each contrast: 81.22% against 78.29% and 80.37% against 77.45%, 2.93 points higher on each iguana contrast (10 images the readout got wrong and the edit got right, 4 the reverse), and 98.33% against 98.33% on the birds. SpeciesNet's standard output scored 12.44%, 12.44% and 3.33%, and its raw top label 27.32%, 27.32% and 17.50%.

Exact genus was harder. The edit raised marine-iguana exact-genus recall from 1.69% to 76.27%, 45 of 59; the unedited model had called marine iguanas land iguanas. Land-iguana recall moved from 98.33% to 95.00% and mockingbird recall from 98.33% to 96.67%. On marine iguanas and mockingbirds the edit stayed below the readout's 89.83% and 100.00%; on land iguanas it was above the readout's 90.00%.

The edit also raised recall on every control species with test photos. On the rodent menu it named 88 of 112 invasive rats as *Rattus*, against 3 unedited and 51 for the readout, an equal-location recall of 81.23% against 2.96% and 46.86%. Recall rose from 0.88% to 97.37% for South American coatis, from 0.00% to 79.49% for European rabbits and from 0.00% to 55.69% for house mice. Pigs stayed weakest, at 37.25% against 72.55% for the readout and 60.29% for SpeciesNet. The rat result covers one direction only, because no native rice-rat test sample was available.

	Crops	Unedited	Check 39	Island edit	SigLIP	SpeciesNet stock	SpeciesNet raw
Native-status balanced accuracy (%), native / look-alike crops							
Land iguana vs green iguana	60 / 57	50.00	51.83	**81.22**	78.29	12.44	27.32
Marine iguana vs green iguana	59 / 57	50.00	51.83	**80.37**	77.45	12.44	27.32
Mockingbird vs brown noddy	60 / 60	70.83	91.67	**98.33**	98.33	3.33	17.50
Exact-genus recall (%)							
Conolophus (land iguana)	60	98.33	98.33	95.00	90.00	0.00	0.00
Amblyrhynchus (marine iguana)	59	1.69	1.69	76.27	89.83	0.00	0.00
Mimus (mockingbird)	60	98.33	90.00	96.67	100.00	0.00	1.67

Table 9: Island edit on the held-out test, 697 citizen-science photo crops. Native directions each come from one county; the 57 green iguanas come from 41 locations and serve both iguana contrasts. Locations count equally within each direction, then the two directions count equally. The list-based arms chose from fixed 11-name menus; SpeciesNet used its own taxonomy, which lacks both native iguana genera and the Galápagos mockingbird species. "Stock" is SpeciesNet's standard ensemble output; "raw" its highest-scoring classifier label. No intervals: each native direction has one location stratum.

6.5 Cost on the camera benchmark

The island edit was also scored on the original 1,611 crops from 45 cameras. It scored 72.14% balanced accuracy, against 64.09% unedited and 81.76% for check 39. Against the unedited model the change was $+8.05$ points (95% whole-camera interval $+3.01$ to $+13.93$; 280 fixes, 110 harms); against check 39 it was -9.62 points (-13.30 to -6.19; 43 fixes, 189 harms). Unusable replies fell from 237 unedited to 3. Each edit is a specialist. The island edit kept the answer-format repair and part of the camera gain, and gave up the rest of the camera-specific improvement.

7 Part three: one model for both jobs

7.1 Two ways to fit both jobs

The third part fits one model to do the camera-trap job and the island job together, in two ways. Both models start from check 39 and learn from the same 2,902 labelled practice crops: 2,095 from the camera practice set and 807 island crops. Neither used selection or test labels in fitting. Both were scored on the 1,611-crop camera benchmark and the 697-crop island test of parts one and two, with the same prompts, decoding and scorers. The Codex research assistant (OpenAI GPT) built and scored both models.

The *direct combined edit* is gradient-free, like the first two parts. One joint ridge fit wrote 27 island-name hint pairs into 54 projector units that check 39 had not changed, chosen for their small original output weights. Its targets were the island reference scores on island practice crops and zero added hint score on camera practice crops, so that camera answers would stay as check 39 gave them. Three projector tensors changed. Every other check-39 weight stayed, including its camera hints and answer-row corrections. Eight fixed candidates crossed four camera-preservation weights with two hint magnitudes. A separate selection set of 194 camera and 426 island crops chose the candidate whose worse change, against check 39 on cameras and against the island edit on islands, was smallest.

The *LoRA fine-tune* is a standard gradient method that learns a small low-rank update to existing weights [15]. It trained rank-8 updates, with alpha 16, to the attention query and value matrices in all 24 language blocks: 540,672 parameters. The vision encoder, projector, token table and output head stayed at check 39. One seeded pass over the practice crops made 363 optimiser updates. The loss predicted each correct common name and its end-of-answer token, with the same domain, camera, class and species weights as the direct fit. The recipe was fixed in advance, with no candidate search. The update was merged into a complete model for testing.

7.2 Results

The LoRA fine-tune held both jobs (Table 10). It scored 87.58% on land against green iguanas, 91.69% on marine against green iguanas, 93.33% on mockingbirds against brown noddies and 88.20% equal-location recall of invasive rats, while scoring 79.22% on the camera benchmark. It beat the direct combined edit on every island task. On the camera benchmark it was 1.48 points below the combined edit (95% whole-camera interval -5.80 to $+2.54$; 88 fixes, 90 harms), a match within noise, and 2.54 points below check 39 (-6.78 to $+1.09$).

Against the island edit of the second part, the fine-tune scored 6.36 and 11.32 points higher on the two iguana contrasts and 6.97 points higher on rat recall, and 5.00 points lower on the birds. It exceeded the SigLIP readout on both iguana contrasts and on rat recall. Most of its iguana gain came from the look-alike side: it named 51 of 57 green iguanas correctly, against 6 for the combined edit, while keeping 51 of 60 land and 55 of 59 marine iguanas native, against 60 and 59. It named 100 of 112 invasive rats as *Rattus*, against 18 for the combined edit.

The direct combined edit held the camera job. It scored 80.70%, 1.06 points below check 39 (-2.59 to $+0.07$). It recovered little of the island job: 55.49% on both iguana contrasts, 90.83% on the birds and 17.70% rat recall, against 81.22%, 80.37%, 98.33% and 81.23% for the island edit. Its iguana scores are low because it named only 6 of 57 green iguanas. The selection set showed the same loss: the chosen candidate's island score there was 39.03 points below the island edit's.

Check 39, a 0.5B general model edited without a training run, scored 81.76% on the camera benchmark, level within noise with BioCLIP 2 [16], an open image model built for species recognition and trained on photos across the tree of life, which scored 84.07% with no adaptation on the same crops and lists ($+2.31$ points for BioCLIP 2, 95% whole-camera interval -1.58 to $+6.38$; 174 fixes, 142 harms; R18, R19). On the island test BioCLIP 2 scored 94.49% and 94.48% on the iguana contrasts, 100.00% on the birds and 83.29% rat recall, having trained on 399 of the 684 iNaturalist test photos. It is a stronger image encoder than the SigLIP tower inside this model, and connecting it to the language model is the next experiment.

7.3 Time to build

Both models were built and scored free on one shared local Mac Studio, in float32 with batch size one. The direct edit took about an hour to build: 8.08 minutes of feature preparation, a 20-second fit and 52.69 minutes to score eight candidates on the selection set. The LoRA fine-tune took 2 h 52 min of training (Table 11). The training ran as 41 attempts separated by 40 memory pauses; each attempt resumed from the last saved update and continued the same pass. The times exclude the earlier work that built check 39, the island edit and their cached image features.

In this study the direct edit is the fast route, and the fine-tune is the route that held both jobs in one model.

	Camera benchmark	Land / green iguana	Marine / green iguana	Mockingbird / brown noddy	Invasive-rat recall
Unedited	64.09	50.00	50.00	70.83	2.96
Check 39 (part one)	**81.76**	51.83	51.83	91.67	13.75
Island edit (part two)	72.14	81.22	80.37	**98.33**	81.23
Direct combined edit	80.70	55.49	55.49	90.83	17.70
LoRA fine-tune	79.22	**87.58**	**91.69**	93.33	**88.20**
SigLIP readout	82.79	78.29	77.45	98.33	46.86
SpeciesNet	–	12.44	12.44	3.33	8.11
BioCLIP 2	84.07	94.49	94.48	100.00	83.29

Table 10: Every edited model on both tests (%). The camera benchmark is equal-camera balanced accuracy on 1,611 crops from 45 cameras. The island columns are native-status balanced accuracy on the 697-crop island test, except invasive-rat recall, which gives the 76 public locations of the 112 rat crops equal weight. Bold marks the best edited model in each column. The direct combined edit and the LoRA fine-tune both start from check 39 and use the same 2,902 practice crops. The last three rows are reference models run as released: the SigLIP readout, SpeciesNet's standard output, which has no camera-benchmark row in this comparison, and BioCLIP 2 with the taxonomic-plus-common-name text of its released camera-trap evaluator.

Stage (minutes)	Direct combined edit	LoRA fine-tune
Feature preparation	8.08	in training
Fitting or training	0.43	171.56
Candidate selection	52.69	0.00
Writing and reloading the full model	0.42	0.55
Build total	61.62	172.11
Camera test	22.79	34.10
Island test	5.37	7.13

Table 11: Running time for each stage on the same shared host. Training time sums the running time of all 41 attempts and excludes the gaps between them. Direct feature preparation includes one failed first attempt and its recovery.

8 Reproduction

The local reproduction package contains the pinned public base, a small delta containing replacement weights, model-loading assets, public-source image manifests, hash-checking fetch code, the frozen prompt and scorer, construction code and saved verification records. The standalone evaluation regenerated the delivered result: 81.7616% equal-camera balanced accuracy and 83.8610% plain accuracy on the same 1,611 crops. Every archived check-39 raw reply and accepted answer matched. A separate checker reparsed the replies and reproduced the paired gain and interval.

Route	Verified outcome
Standalone evaluation	81.7616% balanced accuracy; every archived check-39 reply and accepted answer matched on 1,611 crops.
Public-base delta replay	Exact canonical tensor hashes and tied aliases after saving and reloading the delivered state.
Practice-input recipe rebuild	713/716 tensors byte-identical; three differ. All test replies, tokens, answers and faults match check 39.

Table 12: What the preserved reproduction records establish. The recipe rebuild and the delta replay have different identity guarantees; the first numerical divergence is unresolved. Records R5–R7.

Two reconstruction routes have different guarantees. Replaying the saved replacement weights reloaded through `from_pretrained` with exact equality for every distinct weight array and its tied input/output alias of the delivered state. A separate practice-input builder recomputed the numerical recipe from original public weights and disclosed practice inputs. It matched 713 of 716 canonical tensors byte for byte. Three tensors differed slightly; the first numerical divergence remains unresolved. On the fixed test, the rebuilt model reproduced every check-39 raw reply, token sequence and accepted answer. This establishes answer identity on this set.

The practice builder recomputes the selected mathematical recipe, not the AI editors' search. The replay ran with float32, batch one and Apple silicon MPS. The verification covers this hardware and backend. Source-frame hashes were checked throughout; the recorded fresh URL smoke test exercised one frame from each original dataset, while the

remaining frames were hash-verified local copies. The manifests and fetcher support public-source reconstruction, but a new network download of every frame was not part of that smoke test. Records R5–R7 distinguish these guarantees.

For the island edit, separately written code recomputed every test metric from the saved raw answers of all five arms on the 697 test crops, checked the role assignments and exclusions, and recomputed the camera-benchmark comparison with its whole-camera interval. All values agreed with those reported. The selected weights reload with a recorded hash, and the patch, candidate changes and reload check are kept with the island records (R10–R12).

For the third part, both models were written as complete models and loaded again before testing. The LoRA reload changed exactly the 48 attention query and value matrices and left every other check-39 tensor byte-identical. Separately written code recomputed every camera and island metric for both models from their saved raw answers, with the pairings, intervals and candidate selection, and agreed with every reported value (R13–R17).

9 Discussion

9.1 Why not just use the visual classifier?

For camera-trap animals in its vocabulary, SpeciesNet is the stronger labeller. The first part's value is that an AI editor computed a repair inside another model while keeping its native answer path. It improved the answer model substantially, brought it close to a separate visual readout under short lists, and exposed where that repair stopped helping.

The second part shows where the edited model can do what the classifier cannot. SpeciesNet has no label for the native iguana genera or the Galápagos mockingbird species, and its standard output scored 12.44% and 3.33% on those contrasts. A targeted edit, fitted from 846 practice photos, brought the answer model to 81.22%, 80.37% and 98.33%. A classifier's vocabulary is fixed when it is trained. This model reads its labels from the supplied list, and a computed edit made it choose among them far more accurately.

The third part shows that one small model can hold both jobs. A LoRA fine-tune from check 39 kept the camera benchmark at 79.22% and reached 87.58%, 91.69%, 93.33% and 88.20% on the island tasks. The direct combined edit held the camera job and lost most of the island one. Each single-job direct edit did its own job well. The fine-tune held both jobs in one model, and the direct combined edit was the faster build, about an hour against about three.

The largest individual restoration effect concerned answer format. At the broad two-half level, the gains were close to additive: 17.14 points summed, beside 17.67 combined. Conditional restorations of tied weights answer a different question from allocating shares of that gain.

The unedited SigLIP readout, which used no labelled practice examples, scored at or above the edited model in every matched condition shown: 82.79%/81.76% on the original lists, 56.41%/55.55% on the 11-name list, 43.79%/0.27% on the 64-name list and 13.60%/4.74% in strict new-collection open naming. The edit used 23,139 labelled practice crops plus outputs of that reference. The gain repairs the native answer path; it adds no accuracy beyond the available visual readout in these comparisons. The island edit, fitted from the readout's scores on 846 practice photos, matched the readout on mockingbirds and exceeded it on both iguana contrasts, while staying below it on marine-iguana and mockingbird exact genus. The LoRA fine-tune, trained on practice answers rather than readout scores, exceeded the readout on both iguana contrasts (87.58% and 91.69% against 78.29% and 77.45%) and on rat recall (88.20% against 46.86%), and stayed below it on the birds and the camera benchmark.

9.2 Which parts might carry to other models?

The separation between an image encoder, a connector and a language model is common in open vision-language families. That structure generalises the question: measure what a separate readout can recover from visual features, compare it with native answers under matched inputs, and inspect the connector when useful information is lost along the answer path. The measured gap and repair must still be established in each model. Within this model, the same construction carried from camera-trap lists to island menus by changing the class lists and practice data, which is the part of the method a new place would supply.

The present implementation is specific. It uses register-like feature addresses, a two-layer GELU projector, fitted class maps, a 58-name vocabulary and a tied output table with corrections only through K. A different tokenizer, connector or image layout could require a different numerical construction. The full-list collapse shows why scaling the option vocabulary cannot be assumed to preserve the short-list result.

9.3 A local visual model beneath a reasoning model

One possible applied design is a small visual model fitted to one place by an agent, with a larger model handling questions, decisions and actions around it. The larger model could inspect errors, propose numerical changes and check their effect before using the smaller model's outputs in conservation knowledge work. This case supplies evidence for one part of that design: a frozen small model's delivered decisions improved through direct computed edits, a fixed short-list improvement carried to another collection, an edit aimed at one archipelago's natives taught the model to separate animals a purpose-built classifier lacks labels for, and a fine-tune of a few hours on a local machine kept a camera-trap job and an island job in one small model.

Place-specific adaptation, sustained updating, calibrated uncertainty, human review workload, downstream decisions and the reliability of an acting system were not measured. A classifier may remain the appropriate visual component. An agent's ability to modify that component is a separate capability from proving that the surrounding decision system is useful or safe in practice.

10 Limits and future work

The parts carry different limits. In the first, the editors inherited one another's weights and notes, test scores were inspected at several stages and could influence whether editing continued, and the test cameras had been screened before this session; the camera intervals therefore describe fixed states, not independent editing sessions. The repair is specific to a 58-name practice vocabulary, incompletely covered by labelled examples, and a tied output table corrected through K. The long-list comparison changes label contents and option positions along with length; on full-list truths in positions A–K, both unedited and edited states scored 0.00%, so correction stopping at K does not explain that collapse. In the second part, each native test group comes from one public county, Santa Cruz, so the island contrasts report counts and no confidence intervals. Native and look-alike photos come from different places and may differ in background or photographic style, so the test cannot separate recognition of the animal from those cues. The bird contrast rests on brown noddies alone. Native rice rats could not be tested: on the four photos from the measurement, the island edit kept a native answer for 1 of 4, against 4 of 4 for the unedited model and check 39, and no arm named the genus. SpeciesNet's scores reflect both its image classification and the labels its taxonomy holds, so a missing label does not show whether its image features separate these animals. The images are citizen-science photographs, not camera-trap frames. The same assistant coordinated fitting and evaluation, with file-access restrictions rather than independent blinded teams separating their inputs. In the third part each route is one recipe: the direct edit chose among eight candidates by a fixed rule, and the fine-tune used one fixed configuration with no search.

The next comparisons are the same repair on a larger open model, answer-row corrections beyond K, a direct combined edit that also refits the answer rows, and island natives in camera-trap frames from several islands. These would test greater language capacity, a broader answer vocabulary, whether a direct edit can hold both jobs, and the field setting.

11 Conclusion

Direct computed weight edits repaired this small model's answers under short supplied lists. The original 45-camera score rose from 64.09% to 81.76%. On a new collection the edit closed about 41% of the model's distance to SpeciesNet under an 11-name list, leaving a 39.58-point gap. Open naming made essentially no progress, and the small base model already collapsed under the full 64-name list. In this part the unedited visual readout remained at or above the edited answers.

The same construction, aimed at Galápagos natives, taught the model to separate animals SpeciesNet lacks labels for: both native iguana genera and the Galápagos mockingbird species. On held-out citizen-science photos, mockingbird-versus-brown-noddy balanced accuracy rose from 70.83% to 98.33%, equal to the readout and against 3.33% for SpeciesNet. Land and marine iguanas against green iguanas rose from 50.00% to 81.22% and 80.37%, above the readout's 78.29% and 77.45% and against 12.44% for SpeciesNet. The island edit kept 8.05 of the camera benchmark's points over the unedited model and gave up 9.62 against the general edit.

One model can hold both jobs. Starting from the general edit and the same practice photos, a LoRA fine-tune reached 87.58% and 91.69% on the two iguana contrasts, 93.33% on the birds and 88.20% recall of invasive rats, while scoring 79.22% on the camera benchmark, within noise of the direct combined edit's 80.70%. The direct combined edit held the camera job but not the island one. The direct route was the fast one, about an hour against about three for the fine-tune.

Together the three parts show an AI editor repairing a model's answers for camera traps, aiming the same repair at animals a purpose-built classifier cannot name, and a short fine-tune holding both jobs in one model where a direct combined edit held only one.

A Record and reproduction details

This section provides the end-matter details needed to connect the measurements to their preserved files. The companion `record-index.json` gives the exact file hashes. R1 is `boundary39-table034.json`; R2 is `independent-result049.json`; R3 is `table040.json`; R4 is the Felidae `metrics.json` with its result, split and form reports. R5 is `reproduction.json`; R6 is the practice-construction tensor comparison, score and independent verification; R7 is `model-proof.json`. R8 is the editor brief; R9 is the frozen candidate inventory. For the island edit, R10 is `main-summary019.json`, R11 is `camera-analysis010.json` and R12 is the independent recomputation `independent-result031.json`, with the four-photo rice-rat audit `independent-rice-result032.json`. The selected island candidate is `base-distinctive-060-letters4`. For the third part, R13 is the LoRA summary `summary018.json`, R14 its independent recomputation `independent-result019.json`, R15 the timing arithmetic `comparison-support026.json`, R16 the direct combined summary `summary010.json` and R17 its independent recomputation `independent-result011.json`. For the BioCLIP 2 reference, R18 is `summary002.json` and R19 its independent recomputation `independent003.json`. The selected combined candidate is `joint-rho0.1-mag060`. The records and companion kit are staged locally and are not yet publicly hosted.

Pinned state. The base maker revision is `381d9947148efb1e58a577f451c05705ceec666e`. The delivered state's canonical SHA-256 is `846cdbb58f3ca04c9ccda139d6e2c412921519f11da6e72a9a268480800b9881`. The practice-built state is `c5e548f73914ec42134b409d9442339e5473be014057ce0d9ebe0a9636f58268`. Its head and first projector weight/bias differ from the delivered state; the second projector matrix matches.

Temporal and asset provenance. The Felidae capture cut-off was 19 March 2025; only later captures were eligible. Public release was reported on 3 March 2025 [17]; PyPI records package 4.0.3 on 19 March 2025 [18]. The package release is a dated software record, distinct from the model's unpublished training cut-off. The current neural classifier file matched the earlier 4.0.1a file byte for byte; taxonomy, geofence and label assets were separately pinned. The comparison uses that pinned classifier and separately pinned supporting assets. The OpenAI editor's actual response model was `openai/gpt-6-astra-pro`; the recorded catalogue canonical slug was `openai/gpt-6-astra-pro-20260903`.

Reference-pipeline limitation. The frozen MMStar replies scored 38.53% with official correct/1,500 aggregation versus 37.5% published, and 40.31% with equal-L2 weighting [19]. The maker's aggregation remains unresolved. The unresolved mismatch is retained in the record. This limits claims about reference-pipeline parity without changing the within-study matched comparisons.

Licences and data care. Model and modification notices in the prepared companion kit retain Apache-2.0 terms and upstream attributions. The source image datasets have their own Community Data License Agreement permissive terms. Public human-image exclusions are preserved; the recorded original-test manifest had complete source-image checks and no human-label matches in the Island Conservation portion. The paper does not redistribute full source image collections.

B Complete answer forms and label-position splits

Arm	Condition	Letter	Name	Both agree	Both conflict	Other
Unedited	64 names	0	0	2400	1	1
Check 39	64 names	2152	0	123	127	0
Check 38	64 names	2151	0	123	128	0
Unedited	11 names	0	0	1541	360	0
Check 39	11 names	1900	0	1	0	0
Check 38	11 names	1900	0	1	0	0
Unedited	No list	0	750	0	0	1652
Check 39	No list	0	925	0	0	1477
Check 38	No list	0	955	0	0	1447

Table 13: Felidae raw output forms. The 64-name and no-list conditions have N=2,402; the 11-name condition has N=1,901. These counts include unresolved truth, unlike the accuracy tables. Form bins are exclusive and distinct from correctness.

11-name option	Label	Crops	Unedited	C39	C38	SigLIP	SpeciesNet
A	bobcat	598	45.82	53.90	53.98	51.04	96.93
B	coyote	190	26.41	55.42	56.85	53.76	93.10
C	gray fox	233	92.76	82.48	82.25	34.88	87.98
D	mouse or rat	0	N/A	N/A	N/A	N/A	N/A
E	mule deer	477	0.16	67.58	67.41	68.23	94.29
F	rabbit	63	21.91	35.10	35.10	27.46	99.09
G	raccoon	72	24.03	32.48	32.48	28.30	100.00
H	striped skunk	31	18.06	42.01	42.01	28.47	86.11
I	turkey	27	68.27	79.81	79.81	76.73	100.00
J	unknown bird	52	0.00	61.57	61.57	88.01	100.00
K	unknown squirrel	112	0.00	1.75	0.00	87.10	94.81

Table 14: Every fixed 11-name-list option, equal-location balanced accuracy (%). Bobcat is option A on the 11-name list and option I on the 64-name list. Option D has 46 unresolved crops and zero scorable examples and remains N/A. Source locations overlap between rows; class and option position are confounded.

Acknowledgements and tool use

Claude Fable 5.2, Claude Opus 5.5 and GPT-6 Astra Pro were used as editing tools for the general camera-trap edit. The Codex research assistant (OpenAI GPT) built and evaluated the island edit, the direct combined edit and the LoRA comparison. AI tools also assisted with analysis, figure preparation, manuscript drafting and internal review. They are not authors. Andrew Hill is responsible for the interpretation and manuscript. Procedural scientific-writing and verification guidance was drawn from Scientific Agent Skills [20].

References

[1] Rikiya Takehi, Jayden Teoh, Alex Zhang, and Omar Khattab. Sparks of Weight-Space Programming: LLMs can natively read & write neural network weights. https://rikiyat.github.io/blog/weight-space-programming/, 2026. Author research demonstration.

[2] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. URL https://arxiv.org/abs/2408.03326. arXiv:2408.03326.

[3] 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 (ICCV)*, pages 11975–11986, October 2023.

[4] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671. arXiv:2407.10671.

[5] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191. arXiv:2409.12191.

[6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. URL https://arxiv.org/abs/2305.06500. arXiv:2305.06500.

[7] Google. SpeciesNet: camera-trap species classification software. https://github.com/google/cameratrapai, 2025.

[8] Tomer Gadot, Ștefan Istrate, Hyungwon Kim, Dan Morris, Sara Beery, Tanya Birch, and Jorge Ahumada. To crop or not to crop: Comparing whole-image and cropped classification on a large dataset of camera trap images. *IET Computer Vision*, 18(8):1193–1208, November 2024. ISSN 1751-9640. doi: 10.1049/cvi2.12318. URL http://dx.doi.org/10.1049/cvi2.12318.

[9] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[10] Island Conservation. Island Conservation Camera Traps. https://lila.science/datasets/island-conservation-camera-traps/, 2020. Dataset, Community Data License Agreement, permissive variant.

[11] Felidae Conservation Fund. Felidae Conservation Fund camera-trap dataset. https://lila.science/datasets/felidae-conservation-fund/, 2026. Dataset, Community Data License Agreement, permissive variant.

[12] Alys Granados, Zara McDonald, Katie McPherson, and David Stoner. Unraveling the impact of dog-friendly spaces on urban–wildland pumas and other wildlife. *Wildlife Biology*, August 2024. ISSN 1903-220X. doi: 10.1002/wlb3.01290. URL http://dx.doi.org/10.1002/wlb3.01290.

[13] iNaturalist. iNaturalist research-grade observations. https://www.inaturalist.org/, 2026. Openly licensed (CC0, CC-BY, CC-BY-SA) wild research-grade photographs.

[14] Charles Darwin Foundation. Galapagos species checklists archive. https://datazone.darwinfoundation.org/en/checklist/checklists-archive, 2026.

[15] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[16] Jianyang Gu, Samuel Stevens, Elizabeth G Campolongo, Matthew J Thompson, Net Zhang, Jiaman Wu, Andrei Kopanev, Zheda Mai, Alexander E. White, James Balhoff, Wasila Dahdul, Daniel Rubenstein, Hilmar Lapp, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. BioCLIP 2: Emergent properties from scaling hierarchical contrastive learning, 2025. URL https://arxiv.org/abs/2505.23883. arXiv:2505.23883.

[17] Kyle Wiggers. Google releases speciesnet, an ai model designed to identify wildlife. TechCrunch, https://techcrunch.com/2025/03/03/google-releases-speciesnet-an-ai-model-designed-to-identify-wildlife/, 2025. 3 March.

[18] Google Camera Traps AI team. Speciesnet 4.0.3 package release. PyPI, https://pypi.org/project/speciesnet/4.0.3/, 2025. 19 March.

[19] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024. URL https://arxiv.org/abs/2403.20330. arXiv:2403.20330.

[20] Timothy Kassis, Vinayak Agarwal, Yuhuan He, Darshil Patel, and Aubrey M. Brueckner. Scientific agent skills: A library of procedural knowledge for research agents, 2026. URL https://arxiv.org/abs/2609.00065. arXiv:2609.00065.