

AUTOMATIC IMAGE-LEVEL MORPHOLOGICAL TRAIT ANNOTATION FOR ORGANISMAL IMAGES

Anonymous authors

Paper under double-blind review

ABSTRACT

Morphological traits are physical characteristics of biological organisms that provide vital clues on how organisms interact with their environment. Yet extracting these traits remains a slow, expert-driven process, limiting their use in large-scale ecological studies. A major bottleneck is the absence of high-quality datasets linking biological images to trait-level annotations. In this work, we demonstrate that sparse autoencoders trained on foundation-model features yield monosemantic, spatially grounded neurons that consistently activate on meaningful morphological parts. Leveraging this property, we introduce a trait annotation pipeline that localizes salient regions and uses vision-language prompting to generate interpretable trait descriptions. Using this approach, we construct BIOSCAN-TRAITS, a dataset of 80K trait annotations spanning 19K insect images from BIOSCAN-5M. Human evaluation confirms the biological plausibility of the generated morphological descriptions. When used to fine-tune BioCLIP, a biologically grounded vision-language model, BIOSCAN-TRAITS improves zero-shot species classification on the in-the-wild Insects benchmark, underscoring the value of trait-level supervision for enhancing model generalization.¹

1 INTRODUCTION

The accelerating biodiversity crisis demands rapid advancement in our understanding of ecosystem function and species' responses to environmental change. While taxonomic identification answers the question “what species is this?”, it fails to explain *why* organisms succeed or fail under changing conditions. Morphological traits (the measurable physical characteristics of organisms) provide this critical mechanistic link, predicting with remarkable accuracy how species interact with their environment (Díaz et al., 2016; Kennedy et al., 2020; McGill et al., 2006). Morphological traits can predict species’ ecological niches and functions with up to 85% accuracy (Pigot et al., 2020), offering insights into resource utilization and potential responses to disturbance. Despite their paramount importance, trait data remains trapped in an analog bottleneck: millions of biological specimens and images exist in collections worldwide, but extracting standardized trait measurements requires painstaking manual work by domain experts (Violle et al., 2007), rendering large-scale trait-based ecology virtually impossible.

Measuring even simple characters such as body length or tibia ratio still takes minutes per specimen despite modern digitization techniques (Hardisty et al., 2022). Natural-history institutions curate 3B+ specimens, so a full trait census would consume person-centuries of expert labour (Nelson & Ellis, 2019). Protocols differ by taxon (wing chord for birds, elytral lengths for beetles, sepal length for plants, etc) and this heterogeneity, combined with observer subjectivity, introduces systematic bias that complicates data synthesis (Heberling, 2022). Even when traits are quantified, they often remain in notebooks or image captions, invisible to machine pipelines, leaving a global “trait data desert” that blocks large-scale trait ecology studies.

Automating trait mining pushes ML into a worst-case regime. First, biology’s cross-taxon heterogeneity means the feature manifold warps whenever one moves from, say, angiosperm leaves to wasp antennae; He et al. (2024) lists this taxonomic domain shift as the single largest unsolved barrier to reliable pipelines. Digitized specimens further exhibit uncontrolled pose, preservation artifacts,

¹Code and data will be released on GitHub.

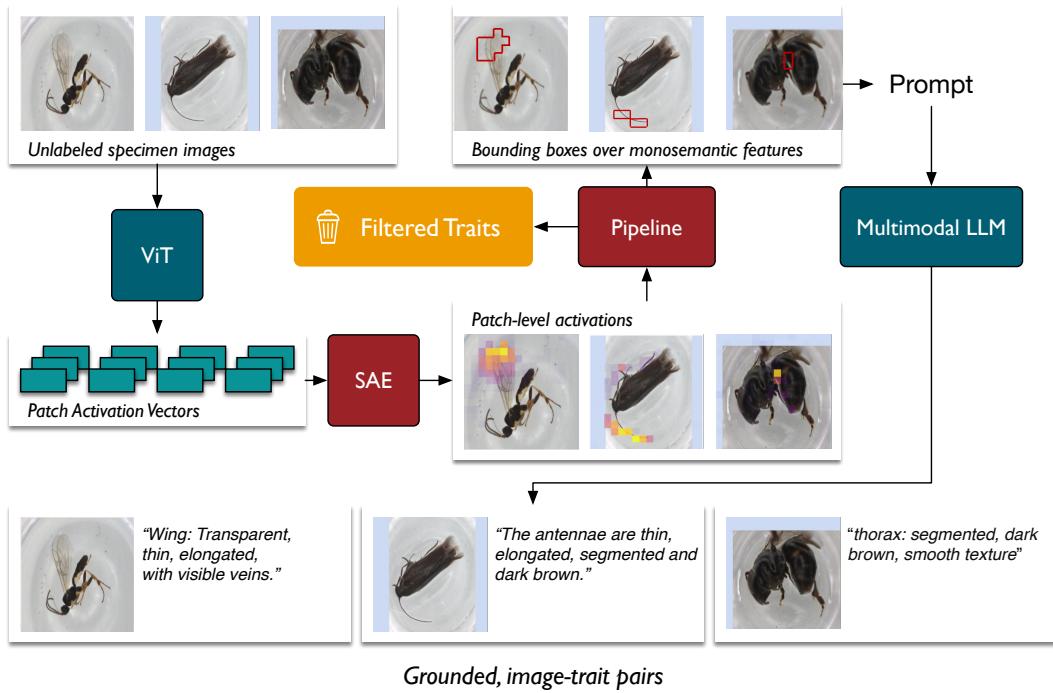


Figure 1: Given an input specimen image, we first compute dense visual representations using an off-the-shelf backbone (e.g., DINOv2). These features are passed through a pre-trained sparse autoencoder (SAE), which identifies high-activation latent units corresponding to semantically meaningful regions (Algorithm 1). We extract the spatial masks associated with these activations and overlay them on the original image to localize trait-relevant boxes. Finally, a multimodal language model (MLLM) is prompted with the annotated image to generate fine-grained morphological trait descriptions. **This results in a large-scale, automatically labeled image-level trait dataset.**

and background clutter, factors that amplify distribution shift and explode the sample complexity demanded by supervised learning. Second, a systematic review of 50+ herbarium-vision papers finds that apparently “simple” tasks (leaf area, margin type) still need bespoke augmentation recipes and hyper-parameter sweeps for every dataset, with no method transferring cleanly across collections (Hussein et al., 2022). Third, even mature semi-automated tools, such as *Inselect* (Hudson et al., 2015) for drawer segmentation, end up handing users a GUI for redrawing boxes; human operators spent 108 seconds per image correcting model outputs. Together, these observations show that standard supervised learning struggles when labels are scarce, morphology is non-stationary, and objects occupy only tiny, variable parts of the frame—precisely the conditions that trait ecology presents.

Our key insight is recognizing that **sparse autoencoders (SAEs) can be used as interpretable part-detectors for trait extraction.** A sparse autoencoder learns, from unlabeled data, a dictionary of latent units that can linearly reconstruct frozen foundation-model embeddings while enforcing two pressures: (i) *sparsity*: only a few units fire for any image, and (ii) *non-negativity*: activations cannot cancel each other. These constraints push each latent unit toward a single, reusable visual cause rather than a mixture of unrelated cues. In practice, training an SAE over pre-trained image features produces units whose activations map back onto tight, spatially coherent regions such as “hind-leg femur band,” “dorsal eye stripe,” or “apical leaf tip.” (see §4.4 for example visualizations) After training, we can (1) isolate just the pixels that define a candidate trait, (2) visually indicate the relevant area, and (3) describe those areas with a vision-language model. To focus on truly diagnostic parts, we introduce a species-contrastive ranking: a unit is valuable when it fires strongly for a target species but remains almost silent for closely related species. High-ranked units, therefore, highlight precisely the salient, fine-scale structures that taxonomists record as traits, making the SAE an ideal front end for our trait-distillation pipeline.

108 We instantiate these ideas in a three-step, concrete, trait-labeling pipeline (Figure 1) and apply it to
 109 the BIOSCAN-5M insect corpus (Gharaee et al., 2024):
 110

- 111 1. We rank SAE units by a species-contrastive score that privileges activations that are strong for a
 focal species yet weak for its congeners.
- 112 2. High-score masks are boxed into tight patches.
- 113 3. Each patch is prompted to a large multimodal large language model (Qwen2.5-VL-72B) with a
 lightweight template.

116 Because the SAE provides locality and taxonomic focus before any language model is consulted, the
 117 MLLM’s task is far easier: “describe this part” rather than “describe the whole scene”, which sharply
 118 reduces hallucinations and background leakage. **While BIOSCAN-5M provides the large-scale,**
 119 **species-labeled data for our experiments, the pipeline itself only requires images with taxonomic**
 120 **labels.** Such supervision that is widely available in many other domains (*e.g.*, iNaturalist (Horn
 121 et al., 2018), TreeOfLife (Stevens et al., 2024), Caltech-UCSD Birds-200-2011 (Wah et al., 2011)).
 122 These resources span plants, birds, fungi, and other taxa, making our approach broadly applicable
 123 for **converting labeled image repositories into rich, interpretable trait annotations.** Using the best-
 124 performing configuration, we label 19K images with 80K morphological trait descriptions (averaging
 125 4.2 traits per image), yielding the BIOSCAN-TRAITS dataset. To evaluate robustness and design
 126 sensitivity, we conduct a comprehensive ablation study, systematically examining how individual
 127 design choices influence the quality of the resulting trait descriptions. As an initial validation,
 128 we fine-tune BioCLIP (Stevens et al., 2024), a biologically grounded vision–language foundation
 129 model on BIOSCAN-TRAITS and observe improved zero-shot species classification on an in-the-wild
 benchmark, highlighting the downstream potential of trait-level supervision.

130 In summary, we contribute (i) a species-contrastive SAE-and-MLLM-based algorithm that turns
 131 unsupervised images into high-fidelity, spatially grounded trait labels and (ii) BIOSCAN-TRAITS, a
 132 large, open, image-trait dataset, and (iii) an initial downstream evaluation showing that fine-tuning
 133 a foundation model on BIOSCAN-TRAITS improves zero-shot species classification on an in-the-
 134 wild benchmark. By annotating traits with a modular pipeline rather than prohibitively expensive
 135 manual efforts, we provide a scalable path for injecting biologically meaningful supervision into
 136 foundation models, bridging the long-standing gap between ecological relevance and machine-
 137 learning practicality.

138 2 RELATED WORK

141 **Sparse Autoencoders.** Sparse autoencoders (SAEs) (Makhzani & Frey, 2013; 2015) have proven
 142 effective for uncovering disentangled and human-interpretable latent factors in high-dimensional
 143 representations (Stevens et al., 2025). Prior work has shown the utility of SAEs to learn improved
 144 image (Makhzani & Frey, 2013; 2015) and word representations (Subramanian et al., 2018). To en-
 145 hance feature disentanglement and interpretability, several architectural variants have been proposed,
 146 including top- k activation mechanisms (Bussmann et al., 2024) and multi-layer Matryoshka encoders
 147 designed to promote hierarchical concept structure (Bussmann et al., 2025). SAEs have also been
 148 applied to the internal activations of transformer-based language models, where they reveal latent
 149 units aligned with semantically meaningful and interpretable concepts (Yun et al., 2021; Bricken
 150 et al., 2023; Gao et al., 2024; Templeton et al., 2024). Recent work demonstrates that, when trained
 151 on embeddings from large pretrained models, SAEs can produce monosemantic features, latent units
 152 that respond consistently to a single semantic concept (Templeton et al., 2024; Pach et al., 2025). In
 153 this work, we extend these insights to the domain of biological vision, using SAEs to construct a
 dataset of fine-grained morphological traits from organismal images.

154 **Morphological Trait Extraction.** Traditionally, morphological analysis has relied on manual
 155 measurements and qualitative trait descriptions—a process that is labor-intensive, time-consuming,
 156 and dependent on domain expertise (Hunt & Pedersen, 2025). While these methods offer valuable
 157 insights, they are inherently difficult to scale to large datasets. Recent approaches have begun to
 158 automate trait extraction by leveraging representation learning. For instance, Hoyal Cuthill et al.
 159 (2019) used a convolutional triplet network to map images into a phenotypic embedding space,
 160 enabling quantitative similarity measures and phenotypic tree reconstruction from purely visual data.
 More recent work has pushed further: deep models that segment relevant image regions (*e.g.*, in
 herbarium scans (Ariouat et al., 2025)) or learn latent representations (*e.g.*, via VAEs (Tsutsumi

et al., 2023)) show that rich morphological information can be captured without hand-engineered features. A key challenge, however, is developing models that remain robust to digitization artifacts and background clutter, while also offering interpretability so that biologists and ecologists can identify which morphological features drive predictions. Our work leverages sparse autoencoders to automatically extract morphological traits in BIOSCAN (Gharaee et al., 2024) specimen images. We posit that such trait-level supervision can enhance the robustness and generalizability of MLLMs for fine-grained taxonomical classification.

3 METHODOLOGY

3.1 BACKGROUND

Sparse autoencoders (SAEs) transform dense representations into sparse encodings, where each unit ideally corresponds to an interpretable latent factor. Given a dense input vector $\mathbf{z} \in \mathbb{R}^d$ from an intermediate layer of a vision transformer, the autoencoder maps \mathbf{z} to a high-dimensional sparse representation $g(\mathbf{z})$, from which \mathbf{z} is subsequently reconstructed. This decomposition reveals structured latent factors while preserving the original information content. We use ReLU autoencoders (Bricken et al., 2023; Templeton et al., 2024) for our experiments.

$$\mathbf{u} = \mathbf{W}_e(\mathbf{z} - \mathbf{b}_d) + \mathbf{b}_e, \quad (1)$$

$$g(\mathbf{z}) = \text{ReLU}(\mathbf{u}), \quad (2)$$

$$\tilde{\mathbf{z}} = \mathbf{W}_d g(\mathbf{z}) + \mathbf{b}_d, \quad (3)$$

where $\mathbf{W}_e \in \mathbb{R}^{n \times d}$, $\mathbf{b}_e \in \mathbb{R}^n$, $\mathbf{W}_d \in \mathbb{R}^{d \times n}$, and $\mathbf{b}_d \in \mathbb{R}^d$. Here, $\mathbf{W}_e \in \mathbb{R}^{n \times d}$ denotes the SAE encoder matrix that maps the dense backbone representation $\mathbf{z} \in \mathbb{R}^d$ to the pre-activation latent vector $\mathbf{u} \in \mathbb{R}^n$, and $\mathbf{W}_d \in \mathbb{R}^{d \times n}$ denotes the decoder matrix that maps the sparse code back to the reconstructed representation $\tilde{\mathbf{z}} \in \mathbb{R}^d$. The encoder and decoder also include bias terms: $\mathbf{b}_e \in \mathbb{R}^n$ and $\mathbf{b}_d \in \mathbb{R}^d$, respectively.

The training objective minimizes the reconstruction error while encouraging sparsity in the latent representation:

$$\mathcal{J}(\phi) = \|\mathbf{z} - \tilde{\mathbf{z}}\|_2^2 + \alpha \mathcal{R}(g(\mathbf{z})), \quad (4)$$

where \mathcal{R} denotes the sparsity regularizer and the sparsity coefficient (α) controls the trade-off between sparsity and reconstruction. We use DINoV2-base (Oquab et al., 2024) as the feature backbone to extract dense visual representations from specimen images (see ablations in Appendix I).

3.2 DATASET GENERATION

We use the high-activation latents (with values above a certain threshold $t_{activation}$) to generate descriptions of salient morphological traits in species images. The trait extraction procedure consists of the following steps:

- 1. Sparse Activation Computation:** For each image in the BIOSCAN-5M dataset annotated at the species level, we compute its sparse latent representation using the trained autoencoder.
- 2. Trait Selection via Activation Thresholding:** From the full set of activated latent features for a given sample, we retain only those whose activation values exceed a predefined threshold (denoted by $t_{activation}$), indicating salient trait expression.
- 3. Taxonomic Trait Aggregation:** We then compute the frequency distribution of activated traits at both the species and genus levels across the dataset.
- 4. Trait Filtering by Prevalence:** Within each taxonomic rank, we retain only those traits whose normalized frequency, computed as the ratio of trait occurrences to the total number of trait activations for the taxon, exceeds a predefined threshold (denoted by t_{freq}). This filtering step mitigates noise and retains consistently expressed traits.
- 5. Salient Trait Identification:** We identify salient morphological traits for a species as the ones expressed in a significantly higher proportion within that species than across its corresponding genus, indicating taxon-specific salience.

Algorithm 1: Salient Trait Extraction from Sparse Autoencoder Activations

Input: Species-labeled dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$
Trained sparse autoencoder f_θ
Activation threshold $t_{\text{activation}}$
Normalized frequency threshold t_{freq}

Output: Set of salient traits $\mathcal{T}_{\text{distinct}}$ for each species

1 Initialize counters C_{species} and C_{genus} as empty maps;
2 **foreach** $(x_i, y_i) \in \mathcal{D}$ **do**
3 $z_i \leftarrow f_\theta(x_i)$; // Sparse latent vector
4 $\mathcal{Z}_i \leftarrow \{z_j \mid z_i[j] > t_{\text{activation}}\}$;
5 **foreach** trait z **do**
6 $C_{\text{species}}[s][z] = \sum_{i:y_i=s} \mathbf{1}[z \in \mathcal{Z}_i]$ $C_{\text{genus}}[g][z] = \sum_{i:\text{genus}(y_i)=g} \mathbf{1}[z \in \mathcal{Z}_i]$
7 **foreach** species s and its genus g **do**
8 **foreach** trait z **do**
9 $f_s(z) \leftarrow \frac{C_{\text{species}}[s][z]}{\sum_{z'} C_{\text{species}}[s][z']}$;
10 $f_g(z) \leftarrow \frac{C_{\text{genus}}[g][z]}{\sum_{z'} C_{\text{genus}}[g][z']}$;
11 Initialize $\mathcal{T}_{\text{distinct}} \leftarrow \{\}$;
12 **foreach** species s with genus g **do**
13 $\mathcal{T}_s \leftarrow \{z \mid f_s(z) > t_{\text{freq}} \wedge f_g(z) > t_{\text{freq}} \wedge f_s(z) > f_g(z)\}$;
14 $\mathcal{T}_{\text{distinct}}[s] \leftarrow \mathcal{T}_s$;
15 **return** $\mathcal{T}_{\text{distinct}}$

Algorithm 1 illustrates the procedure for selection of salient traits in detail. Given these traits, we prompt multimodal language models to query the morphological trait descriptions (Figure 1). Prompt templates are provided in Appendix C, and dataset statistics are in Table D.5. Additional dataset examples are shown in Appendix F.

4 EXPERIMENTS

4.1 SPARSE AUTOENCODER TRAINING

We use the BIOSCAN-5M dataset (Gharaee et al., 2024) for our experiments. BIOSCAN-5M is a comprehensive dataset of insect specimens with multiple modalities, including images, DNA barcodes, taxonomical, geographical, and size information. It contains insect images annotated at different levels of the taxonomic hierarchy, with 9.2% of the samples annotated at species-level. While BIOSCAN-5M provides the large-scale, species-labeled data used in our experiments, the pipeline itself only requires collections of images accompanied by taxonomic labels, a type of supervision that is widely available in many other domains (*e.g.*, iNaturalist (Horn et al., 2018) and TreeOfLife (Stevens et al., 2024)). Such resources span plants, birds, fungi, and other taxa, making our approach general and scalable for converting species and genus-labeled biological image repositories into rich, interpretable trait-level annotations.

We train the sparse autoencoder on the entire set of images in BIOSCAN-5M, while the trait generation pipeline uses the subset with species-level labels. The complete hyperparameter setup is given in Table D.4 in the Appendix.

4.2 COMPARISON WITH GRAD-CAM

We compare our pipeline to using traditional feature visualization approaches like Grad-CAM (Selvaraju et al., 2017) for obtaining saliency maps and then forwarding to the MLLM for trait generation. While Grad-CAM can highlight salient regions for a given class label, it lacks trait-level

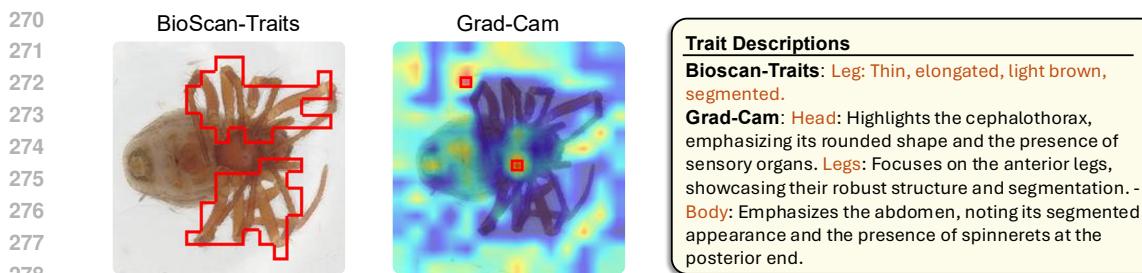


Figure 2: Comparison of trait localization for *Thymoites guanicae*. **BioScan-Traits (left)** generates interpretable trait descriptions that are tied to clear, specific anatomical structures. In contrast, **Grad-Cam (center)** produces diffuse heatmaps that highlight broad body areas without species-level disentanglement.

Table 1: Incorporating latent-specific patches significantly improves the quality of trait descriptions. Including multiple images in the prompt encourages MLLMs to focus on the traits common across all images, at the cost of more tokens per query. Using multiple images with SAE-extracted bounding boxes leads to improved precision, as better ratings indicate. We report both raw and mean-normalized ratings. The experimental setup uses Qwen2.5-VL-72B as MLLM, a normalized frequency threshold of (t_{freq}) = $3e-3$, and 1,000 input images.

Method	# Images	# Tokens	# Images	# Traits	Avg. Raw Rating	Avg. Rating
MLLM	1	413	—	—	3.01	3.00
MLLM	3	940	—	—	3.12	3.15
MLLM + SAE	1	411	460	9,435	3.92	3.84
MLLM + SAE	3	1,072	460	7,897	4.01	3.91

disentanglement, *i.e.*, its heatmaps typically blend multiple anatomical cues, making it difficult for an MLLM to generate precise, interpretable trait descriptions. Moreover, Grad-CAM activations are not species-discriminative, often capturing features shared across related taxa (genus or family level), whereas our SAE-based approach explicitly isolates species-specific, monosemantic neurons tied to fine-grained traits.

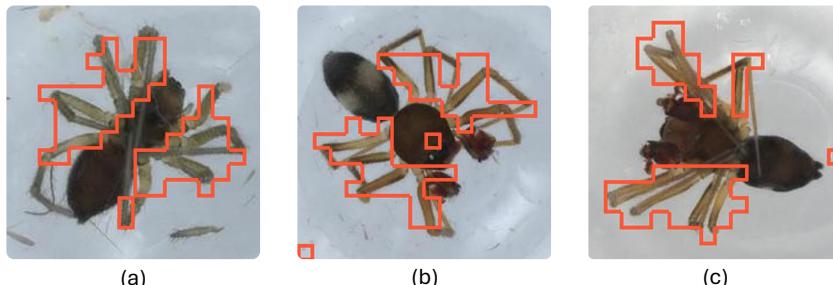
4.3 DATASET ABLATIONS

We conduct a series of ablation studies to evaluate the impact of key design choices on the accuracy and plausibility of trait annotations. For each configuration, we randomly sample 30 trait descriptions and evaluate them using a five-point rubric. Three domain experts independently rated the samples. We apply per-rater mean normalization to ratings, rescaling each annotator’s scores so that their personal mean equals the global mean (Riley et al., 2024; Kirk et al., 2024). This ensures that differences in individual scale usage (*e.g.*, consistently harsh or lenient raters) do not skew aggregated results. The evaluation rubric is given in Appendix E.

Comparison with MLLM-only baseline. As a baseline, we prompt a multimodal large language model (MLLM) with just the specimen image(s) without the trait localization and request a description of salient morphological traits (Figure 3). We compare this to our SAE-guided trait extraction pipeline, which localizes trait-relevant regions via sparse latent activations (Table 1). Incorporating latent-specific patches leads to a substantial improvement in description quality: the average human rating increases from 3.15 to 3.91 in the multi-image setting, highlighting the benefits of spatial grounding provided by the sparse autoencoder for fine-grained trait extraction.

Multiple vs. Single Image per Latent. We investigate the effect of varying the number of input images on trait quality by comparing single-image against 3-image prompts to the multimodal language model (Table 1). Providing multiple images of the same species encourages the model to focus on consistent, shared morphological features while suppressing spurious or image-specific

324
325
326
327
328
329
330
331
332
333



(a) (b) (c)

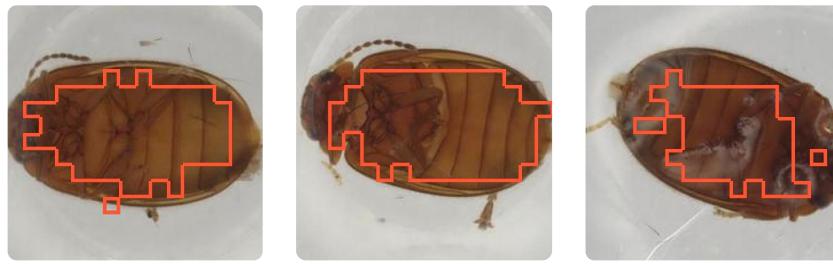
MLLM: Common Morphological Traits Across All Three Images: 1. Head: - Shape: Rounded. - Color: Dark brown or black. - Texture: Appears smooth with no distinct markings. 2. Thorax: - Shape: Oval and slightly elongated. - Color: Lighter than the head, ranging from light brown to beige. - Texture: Smooth, with no visible segmentation. 3. Abdomen: - Shape: Rounded and bulbous. - Color: Dark brown or black, similar to the head. - Texture: Smooth, with no distinct markings. 4. Legs: - Number: Eight legs are visible in all images. - Shape: Thin and segmented. - Color: Light brown, matching the thorax.

MLLM + SAE: Leg: Thin, elongated, light brown, segmented.

340
341
342
343
344

Figure 3: Comparison of salient morphological trait description generation using a just MLLM vs. MLLM + SAE ($t_{\text{freq}} = 1e - 2$) for *Agyneta straminicola*. Each red box highlights a region selected by SAE neurons with high activation, indicating regions used for prompting the MLLM + SAE. The use of SAE helps MLLMs focus on salient morphological traits rather than general descriptions of all body parts.

345
346
347
348
349
350
351
352
353
354
355



(a) (b) (c)

(a) Single image: The highlighted area appears to encompass the main body (thorax and abdomen) of the insect. **Observable Morphological Features:** - Shape: The body has an elongated oval shape. - Size: It occupies a significant portion of the image but cannot be quantified without scale. - Color: The body is primarily brown with some darker and lighter shading, suggesting segmentation or texture. - Texture: The surface appears somewhat smooth but with fine details that could indicate hairs or scales, though this is not entirely clear. - Distinct Markings: There are no prominent markings visible within the highlighted area; the coloration is relatively uniform.

(a, b, c) Multiple images: Body: Brown, oval-shaped, segmented exoskeleton.

362
363
364
365
366

Figure 4: Comparison of salient morphological trait description generation using a single image vs. three images for *Contacyphon ochraceus*. Each red box highlights a region selected by SAE neurons with high activation, indicating regions used for prompting the MLLM + SAE. The use of multiple images yields a concise and taxonomically meaningful output, isolating traits with clearer morphological grounding.

367
368

369 traits. This consensus-driven trait extraction leads to improved precision, as reflected by an increase 370 in the average human rating from 3.84 to 3.91, albeit at the cost of higher token usage per query. A 371 similar trend holds for the MLLM-only baseline.

372
373
374
375
376
377

Additionally, we do a qualitative analysis of the morphological trait descriptions generated by both approaches (Figure 4). Using a single image often leads to trait descriptions that overfit to idiosyncratic visual details of that instance, frequently summarizing multiple anatomical regions, as seen in the example where both the legs and abdomen are described together. This broad coverage can dilute trait precision and obscure what is taxonomically distinctive. In contrast, prompting the model both with multiple images and latent-specific regions encourages it to extract traits that are visually consistent across specimens. This consensus constraint filters out incidental details and leads

378
 379. **Table 2:** SAEs often trade off between reconstruction error (MSE) and sparsity (L_0). We investigate
 380. the effect of choosing between different balances of these errors. We find that lower sparsity performs
 381. better for both values of frequency threshold (t_{freq}). A lower value of the sparsity coefficient (α) leads
 382. to lower MSE and thus better reconstruction. It improves the coverage of latents, leading to better
 383. recall. The experimental setup uses an input dataset of 1,000 images.

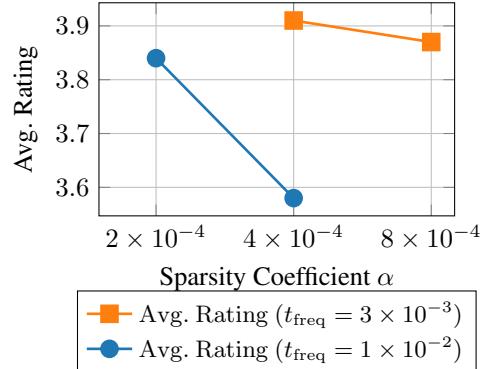
Method	α	t_{freq}	SAE MSE	SAE L_0	# Images	# Traits	Avg. Rating
MLLM+SAE	2e-4	1e-2	8.8e-3	1,081.1	60	60	3.84
MLLM+SAE	<u>4e-4</u>	<u>3e-3</u>	2.7e-2	690.4	460	7,897	<u>3.91</u>
MLLM+SAE	4e-4	1e-2	2.7e-2	690.4	20	20	3.58
MLLM+SAE	8e-4	3e-3	5.4e-2	242.2	458	3,060	3.87

390
 391 to more focused, high-precision descriptions (e.g., isolating just the leg features). As shown in Figure
 392 4, the multi-image setup yields a concise and taxonomically meaningful output, isolating traits with
 393 clearer morphological grounding and higher inter-image agreement.
 394

395
 396 **SAE Quality.** We investigate the sparse autoencoder’s inherent tradeoff between reconstruction error
 397 and sparsity and its downstream impact on morphological trait generation (Table 2). Specifically,
 398 we compare performance across varying values of the sparsity regularization coefficient (α), which
 399 controls the L_0 -sparsity of the latent representation. We observe that lower sparsity (i.e., smaller
 400 α , larger L_0) consistently yields better performance across both values of the normalized frequency
 401 threshold t_{freq} (Figure 5). This setting results in lower mean squared error (MSE), indicating improved
 402 input reconstruction. Importantly, reduced sparsity increases the number of activated latents per
 403 image, thereby improving trait coverage and recall in the final description set.
 404

405
 406 **Table 3:** Effect of normalized frequency threshold
 407 (t_{freq}) on trait selection. We analyze how varying
 408 t_{freq} , which controls the minimum intra-species
 409 normalized frequency required to retain a latent
 410 feature, impacts trait extraction. Lower thresholds
 411 include all activated traits, while higher thresholds
 412 restrict output to only the most consistently ex-
 413 pressed traits. Increasing t_{freq} improves precision
 414 but reduces the number of extracted traits, reflect-
 415 ing a trade-off between coverage and specificity.
 416

Method	t_{freq}	# Images	# Traits
MLLM+SAE	3e-3	460	7,897
MLLM+SAE	6e-3	322	785
MLLM+SAE	1e-2	20	20



421
 422 **Figure 5:** Variation of rating with different lev-
 423 els of SAE sparsity. A lower level of sparsity
 424 performs better for both values of frequency
 425 threshold t_{freq} .
 426

427
 428 **SAE Filtering.** We analyze the effect of the normalized frequency threshold t_{freq} on the trait through-
 429 put using 1,000 input images and sparsity coefficient (α) = $4e - 4$. We observe that increasing t_{freq}
 430 leads to a progressive reduction in the number of retained latent features (Table 3). This results in
 431 the selection of only the more consistently activated latents across a taxon, effectively narrowing the
 432 subset of input images that contribute to trait descriptions. Thus, t_{freq} acts as a precision-recall knob:
 433 lower values yield broader trait coverage but more noise, while higher values emphasize dominant,
 434 taxonomically stable traits.
 435

436
 437 **MLLM Quality.** We compare Qwen2.5-VL-7B and Qwen2.5-VL-72B (Wang et al., 2024) for trait
 438 generation from latent-indexed patches. The larger 72B model yields higher human evaluation scores
 439 and better spatial grounding, avoiding false positive traits; see Appendix B for details.
 440

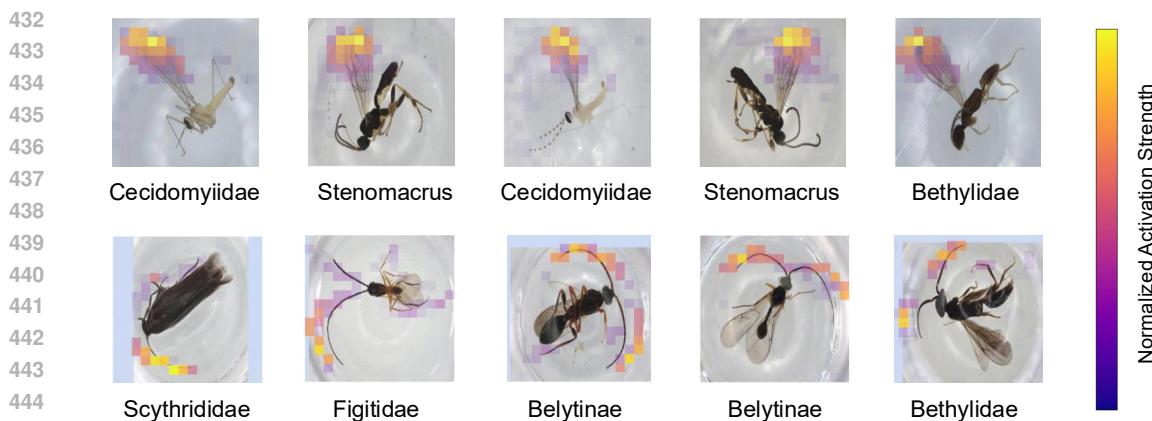


Figure 6: Neurons 4852 and 13860 in SAE get activated at the wings and antennae of insects, respectively. The labels denote the highest annotated taxonomic level. Additional examples are shown in Appendix G.

4.4 NEURON ACTIVATION ANALYSIS

We analyze the top-activating neurons (or latent dimensions) in the SAE to investigate whether they correspond to meaningful morphological traits. Representative examples are shown in Figure 6. Notably, neuron 4852 consistently activates on insect wings, while neuron 13860 responds to antennae, suggesting that specific neurons in the sparse representation are aligned with semantically coherent, interpretable, and biologically plausible traits.

4.5 FINE-TUNING WITH TRAIT SUPERVISION

Table 4: Zero-shot species classification accuracy (%) on the Insects (Ullah et al., 2022) benchmark. Incorporating trait-level supervision yields clear gains over species-only fine-tuning.

Model	Accuracy (%)
BioCLIP (Stevens et al., 2024)	34.8
BioCLIP + species-level fine-tuning (BIOSCAN-TRAITS)	39.6
BioCLIP + trait-level fine-tuning (BIOSCAN-TRAITS)	39.9

To assess the utility of our morphological trait description dataset, we fine-tuned BioCLIP (Stevens et al., 2024), a biologically grounded vision–language foundation model on this dataset. When evaluated on Insects (Ullah et al., 2022), a volunteer-labeled, in-the-wild benchmark, this yielded a significant gain in zero-shot species classification over the pre-trained model. This provides initial evidence that trait-level supervision supports better generalization, underscoring the potential of our dataset for training biologically grounded foundation models.

Notably, sparse autoencoders disentangle foreground from background by activating distinct neuron subsets. By aggregating consistent traits across multiple images per species, our pipeline further improves robustness to real-world noise. As a result, models fine-tuned on SAE-derived trait descriptions generalize more effectively to challenging, in-the-wild imagery.

4.6 COST-OF-USE ANALYSIS

We next quantify the efficiency and cost of our pipeline and examine how well the SAE-guided prompting strategy transfers across different MLLMs. Table 5 reports runtime and throughput on the BIOSCAN-TRAITS workload using two NVIDIA A100 80GB GPUs. The SAE introduces only a small overhead: DINOv2 activation computation and the SAE forward pass together take 7.26 ms per image, whereas MLLM inference (conditioning on three SAE-selected patches per image) dominates the budget at 4.62 s per annotation.

486 Table 5: Runtime and throughput of the proposed pipeline, measured on two NVIDIA A100 80GB
 487 GPUs. Times are averaged over the BIOSCAN-TRAITS workload.

489 Task	490 Time	491 Remarks
491 Activation computation (1 image)	492 2.74 ms	DINOv2 backbone
492 SAE forward (1 image)	493 4.53 ms	Sparse Autoencoder
493 Total preprocessing (1 image)	494 7.26 ms	Feature extraction + SAE
494 MLLM inference (3 images / annotation)	495 4.62 s	Qwen2.5-VL-72B
495 Throughput (2 NVIDIA A100 80GB GPUs)	496 208.9 annotations/h/GPU	

496 Table 6: Cost-of-use analysis for generating trait annotations with Qwen2.5-VL-72B (together.ai) and
 497 GPT-5-mini APIs, reported as average cost per annotation and extrapolated total cost for processing
 498 100K images. The cost is averaged over the BIOSCAN-TRAITS workload, with 1,072 input tokens
 499 and 250 output tokens per annotation.

501 Model	502 Cost per annotation	503 Approx. cost for 100K images
502 Qwen2.5-VL-72B API (together.ai)	503 $\$4.1 \times 10^{-3}$ /annotation	504 \$410
503 GPT-5-mini API	504 $\$8 \times 10^{-4}$ /annotation	505 \$80

506 Table 6 summarizes the cost-of-use when calling Qwen2.5-VL-72B and GPT-5-mini via public
 507 APIs. Closed models such as GPT-5-mini offer stronger performance at lower marginal API cost. In
 508 practice, the open-source Qwen2.5-VL-72B can be hosted in-house, shifting cost from per-call API
 509 pricing to amortized compute, and allowing users with data-governance constraints to keep images
 510 on-premises.

512 CONCLUSION

514 We present a novel pipeline for distilling morphological traits into high-fidelity, natural language
 515 descriptions by leveraging sparse autoencoders and multimodal language models. Applied to the
 516 BIOSCAN-5M dataset, our method produces a large-scale corpus of over 80K trait descriptions
 517 across 19K insect images, constituting one of the first datasets to provide structured, interpretable
 518 trait-level supervision at scale. This dataset serves as a valuable resource for training and evaluating
 519 biological foundation models, particularly in the context of fine-grained taxonomic classification.
 520 Through extensive analysis, we examine the impact of key design factors, including the use of
 521 multiple images for trait verbalization, trait frequency thresholds, sparsity levels in the autoencoder,
 522 and the choice of MLLM backbone, on the precision and accuracy of generated traits. Integrating
 523 trait-level supervision improves generalization in downstream tasks such as fine-grained species
 524 classification, underscoring the utility of our proposed pipeline-generated datasets for biologically
 525 grounded learning. Looking forward, we aim to extend this pipeline to construct large-scale datasets
 526 across diverse biological domains and across multiple taxonomic levels, enabling domain-specific
 527 vision-language models with improved robustness, interpretability, and ecological relevance for
 528 large-scale biodiversity applications.

529 LIMITATIONS

531 We assume that the dense features from the backbone image foundation model encode morphology-
 532 relevant signals. If these representations are biased toward generic visual concepts, important
 533 biological traits may be underrepresented. The SAE discovers latent factors that are spatially
 534 and semantically coherent, but some latents might correspond to multiple co-occurring traits (*e.g.*,
 535 “elongated + thin”) (Chanin et al., 2025). This can make it difficult to disentangle fine-grained trait
 536 attributes or compositional traits. Trait descriptions generated with smaller MLLMs like Qwen-
537 2.5-VL-7B are susceptible to hallucination, particularly when prompted with noisy or background-
538 dominated patches. Also, evaluating trait correctness at scale remains a challenge due to the absence
 539 of ground-truth morphological trait annotations. Please refer to Section A for a more thorough
 discussion of SAEs.

540 ETHICS STATEMENT

541
542 This work advances global biodiversity conservation by introducing a scalable trait annotation pipeline
543 for generating image-to-trait datasets, which can support the development of biologically grounded
544 foundation models. Such models have the potential to improve species recognition, facilitate under-
545 standing of evolutionary patterns, and inform conservation strategies in the context of climate
546 change. By reducing reliance on expert-curated annotations, our approach democratizes access
547 to morphological data and empowers under-resourced institutions and citizen science efforts with
548 automated analysis tools. However, errors in trait interpretation, such as those arising from hallucina-
549 tion or domain shift, may propagate into downstream applications, including species classification
550 and conservation decision-making. It is therefore essential that these tools be deployed in close
551 collaboration with domain experts to ensure reliability and accuracy.

552 REPRODUCIBILITY STATEMENT

553 Our code is available at [https://anonymous.4open.science/r/morph_trait_](https://anonymous.4open.science/r/morph_trait_annotation-8016)
554 annotation-8016, and the dataset can be accessed at [https://www.kaggle.com/](https://www.kaggle.com/datasets/anonymoussub345/morph-trait-annot-bioscan-80k/)
555 datasets/anonymoussub345/morph-trait-annot-bioscan-80k/. The hyperpa-
556 rameter settings are listed in Table D.4, the trait generation pipeline is outlined in Algorithm 1,
557 and prompt templates are provided in Appendix C. All experiments were performed on NVIDIA
558 H100 GPUs.

559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593

594 REFERENCES
595

- 596 Hanane Ariouat, Youcef Sklab, Edi Prifti, Jean-Daniel Zucker, and Eric Chemin. Enhancing plant
597 morphological trait identification in herbarium collections through deep learning-based seg-
598 mentation. *Applications in Plant Sciences*, pp. e70000, 2025. URL <https://bsapubs.onlinelibrary.wiley.com/doi/full/10.1002/aps3.70000?af=R>.
- 599
- 600 Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick
601 Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec,
602 Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina
603 Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and
604 Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary
605 learning. *Transformer Circuits Thread*, 2023. <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.
- 606
- 607 Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. In *NeurIPS
608 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024. URL <https://openreview.net/forum?id=d4dpOCqybL>.
- 609
- 610 Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features
611 with matryoshka sparse autoencoders. In *Forty-second International Conference on Machine
612 Learning*, 2025. URL <https://openreview.net/forum?id=m25T5rAy43>.
- 613
- 614 Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and
615 Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF
616 International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October
617 10-17, 2021*, pp. 9630–9640. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00951. URL <https://doi.org/10.1109/ICCV48922.2021.00951>.
- 618
- 619 David Chanin, Tomáš Dulka, and Adrià Garriga-Alonso. Feature hedging: Correlated features break
620 narrow sparse autoencoders. *arXiv preprint arXiv:2505.11756*, 2025.
- 621
- 622 Sandra Díaz, Jens Kattge, Johannes HC Cornelissen, Ian J Wright, Sandra Lavorel, Stéphane Dray,
623 Björn Reu, Michael Kleyer, Christian Wirth, I Colin Prentice, et al. The global spectrum of plant
624 form and function. *Nature*, 529(7585):167–171, 2016. URL <https://www.nature.com/articles/nature16489>.
- 625
- 626 Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya
627 Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint
628 arXiv:2406.04093*, 2024. URL <https://arxiv.org/abs/2406.04093>.
- 629
- 630 Zahra Gharaee, Scott C. Lowe, ZeMing Gong, Pablo Millan Arias, Nicholas Pellegrino, Austin T.
631 Wang, Joakim Bruslund Haurum, Iuliia Eyrayi, Lila Kari, Dirk Steinke, Graham W. Taylor, Paul W.
632 Fieguth, and Angel X. Chang. BIOSCAN-5M: A multimodal dataset for insect biodiversity. In Amir
633 Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomeczak, and
634 Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference
635 on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada,
636 December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/3fdbb472813041c9ecef04c20c2b1e5a-Abstract-Datasets_and_Benchmarks_Track.html.
- 637
- 638
- 639 Alex R. Hardisty, Paul Brack, Carole A. Goble, Laurence Livermore, Ben Scott, Quentin Groom,
640 Stuart Owen, and Stian Soiland-Reyes. The specimen data refinery: A canonical workflow
641 framework and FAIR digital object approach to speeding up digital mobilisation of natural history
642 collections. *Data Intell.*, 4(2):320–341, 2022. doi: 10.1162/DINT_A_00134. URL https://doi.org/10.1162/dint_a_00134.
- 643
- 644 Yichen He, James M Mulqueeny, Emily C Watt, Arianna Salili-James, Nicole S Barber, Marco Ca-
645 maiti, Eloise SE Hunt, Oliver Kippax-Chui, Andrew Knapp, Agnese Lanzetti, et al. Opportunities
646 and challenges in applying ai to evolutionary morphology. *Integrative Organismal Biology*, 6(1):
647 obae036, 2024. URL <https://academic.oup.com/iob/article/6/1/obae036/7769702>.

- 648 J Mason Heberling. Herbaria as big data sources of plant traits. *International Journal of Plant
649 Sciences*, 183(2):87–118, 2022. URL [https://www.journals.uchicago.edu/doi/
650 full/10.1086/717623](https://www.journals.uchicago.edu/doi/full/10.1086/717623).
- 651
652 Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard,
653 Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist species classification
654 and detection dataset. In *2018 IEEE Conference on Computer Vision and Pattern Recog-
655 nition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 8769–8778. Com-
656 puter Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00914.
657 URL http://openaccess.thecvf.com/content_cvpr_2018/html/Van_Horn_The_INaturalist_Species_CVPR_2018_paper.html.
- 658
659 Jennifer F Hoyal Cuthill, Nicholas Guttenberg, Sophie Ledger, Robyn Crowther, and Blanca Huer-
660 tas. Deep learning on butterfly phenotypes tests evolution’s oldest mathematical model. *Sci-
661 ence advances*, 5(8):eaaw4967, 2019. URL [https://www.science.org/doi/10.1126/
662 sciadv.aaw4967](https://www.science.org/doi/10.1126/sciadv.aaw4967).
- 663 Lawrence N Hudson, Vladimir Blagoderov, Alice Heaton, Pieter Holtzhausen, Laurence Livermore,
664 Benjamin W Price, Stéfan van der Walt, and Vincent S Smith. Inselect: automating the digitization
665 of natural history collections. *PLoS one*, 10(11):e0143402, 2015. URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0143402>.
- 666
667 Roberta Hunt and Kim Steenstrup Pedersen. The phantom of the elytra–phylogenetic trait ex-
668 traction from images of rove beetles using deep learning—is the mask enough? *arXiv preprint
669 arXiv:2502.04541*, 2025. URL <https://arxiv.org/abs/2502.04541>.
- 670
671 Burhan Rashid Hussein, Owais Ahmed Malik, Wee-Hong Ong, and Johan Willem Frederik Slik. Ap-
672 plications of computer vision and machine learning techniques for digitized herbarium specimens:
673 A systematic literature review. *Ecological Informatics*, 69:101641, 2022. ISSN 1574-9541. doi:
674 <https://doi.org/10.1016/j.ecoinf.2022.101641>. URL <https://www.sciencedirect.com/science/article/pii/S1574954122000905>.
- 675
676 Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda.
677 Are sparse autoencoders useful? a case study in sparse probing. In *Forty-second International
678 Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=rNfzT8YkgO>.
- 679
680 Jonathan D Kennedy, Petter Z Marki, Jon Fjeldså, and Carsten Rahbek. The association between mor-
681 phological and ecological characters across a global passerine radiation. *Journal of Animal Ecology*,
682 89(4):1094–1108, 2020. URL <https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/1365-2656.13169>.
- 683
684 Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew M. Bean, Katerina Margatina,
685 Rafael Mosquera Gómez, Juan Ciro, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott
686 Hale. The PRISM alignment dataset: What participatory, representative and individualised human
687 feedback reveals about the subjective and multicultural alignment of large language models. In Amir
688 Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and
689 Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference
690 on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada,
691 December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/be2e1b68b44f2419e19f6c35a1b8cf35-Abstract-Datasets_and_Benchmarks_Track.html.
- 692
693 Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.
694 URL <https://arxiv.org/abs/1312.5663>.
- 695
696 Alireza Makhzani and Brendan J. Frey. Winner-take-all autoencoders. In Corinna Cortes,
697 Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (eds.), *Ad-
698 vances in Neural Information Processing Systems 28: Annual Conference on Neural Infor-
699 mation Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp.
700 2791–2799, 2015. URL <https://proceedings.neurips.cc/paper/2015/hash/5129a5ddcd0dcd755232baa04c231698-Abstract.html>.

- 702 Brian J McGill, Brian J Enquist, Evan Weiher, and Mark Westoby. Rebuilding community ecology
 703 from functional traits. *Trends in ecology & evolution*, 21(4):178–185, 2006. URL <https://www.sciencedirect.com/science/article/pii/S0169534706000334>.
- 704
- 705 Gil Nelson and Shari Ellis. The history and impact of digitization and digital data mobilization on
 706 biodiversity research. *Philosophical Transactions of the Royal Society B*, 374(1763):20170391,
 707 2019. URL <https://royalsocietypublishing.org/doi/10.1098/rstb.2017.0391>.
- 708
- 709
- 710 Maxime Oquab, Timothée Dariset, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov,
 711 Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas
 712 Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael
 713 Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut,
 714 Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision.
 715 *Trans. Mach. Learn. Res.*, 2024, 2024. URL <https://openreview.net/forum?id=a68SUT6zFt>.
- 716
- 717 Mateusz Pach, Shyamgopal Karthik, Quentin Bouriot, Serge Belongie, and Zeynep Akata.
 718 Sparse autoencoders learn monosemantic features in vision-language models. *arXiv preprint*
 719 *arXiv:2504.02821*, 2025. URL <https://arxiv.org/abs/2504.02821>.
- 720
- 721 Alex L Pigot, Catherine Sheard, Eliot T Miller, Tom P Bregman, Benjamin G Freeman, Uri Roll,
 722 Nathalie Seddon, Christopher H Trisos, Brian C Weeks, and Joseph A Tobias. Macroevolu-
 723 tionary convergence connects morphological form to ecological function in birds. *Nature*
 724 *Ecology & Evolution*, 4(2):230–239, 2020. URL <https://www.nature.com/articles/s41559-019-1070-4>.
- 725
- 726 Parker Riley, Daniel Deutsch, George F. Foster, Viresh Ratnakar, Ali Dabirmoghaddam, and Markus
 727 Freitag. Finding replicable human evaluations via stable ranking probability. In Kevin Duh, Helena
 728 Gómez-Adorno, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North Amer-
 729 ican Chapter of the Association for Computational Linguistics: Human Language Technologies
 730 (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 4908–4919.
 731 Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.275.
 732 URL <https://doi.org/10.18653/v1/2024.nacl-long.275>.
- 733
- 734 Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh,
 735 and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localiza-
 736 tion. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October
 737 22-29, 2017*, pp. 618–626. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.74. URL
 738 <https://doi.org/10.1109/ICCV.2017.74>.
- 739
- 740 Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song,
 741 David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf,
 742 Wei-Lun Chao, and Yu Su. Bioclip: A vision foundation model for the tree of life. In *2024
 743 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA,
 744 USA, June 17-21, 2024*, pp. 19412–19424. Computer Vision Foundation / IEEE Computer
 745 Society, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Stevens_BioCLIP_A_Vision_Foundation_Model_for_the_Tree_of_Life_CVPR_2024_paper.html.
- 746
- 747 Samuel Stevens, Wei-Lun Chao, Tanya Berger-Wolf, and Yu Su. Sparse autoencoders for scientifically
 748 rigorous interpretation of vision models. *arXiv preprint arXiv:2502.06755*, 2025. URL <https://arxiv.org/abs/2502.06755>.
- 749
- 750 Anant Subramanian, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard H. Hovy.
 751 SPINE: sparse interpretable neural embeddings. In Sheila A. McIlraith and Kilian Q. Weinberger
 752 (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18),
 753 the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium
 754 on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA,
 755 February 2-7, 2018*, pp. 4921–4928. AAAI Press, 2018. doi: 10.1609/AAAI.V32I1.11935. URL
<https://doi.org/10.1609/aaai.v32i1.11935>.

756 Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen,
 757 Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L
 758 Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers,
 759 Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan.
 760 Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Trans-*
 761 *former Circuits Thread*, 2024. URL <https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html>.

763 Masato Tsutsumi, Nen Saito, Daisuke Koyabu, and Chikara Furusawa. A deep learning approach for
 764 morphological feature extraction based on variational auto-encoder: an application to mandible
 765 shape. *NPJ systems biology and applications*, 9(1):30, 2023. URL <https://www.nature.com/articles/s41540-023-00293-6>.

766 Ihsan Ullah, Dustin Carrión-Ojeda, Sergio Escalera, Isabelle Guyon, Mike Huisman, Felix
 767 Mohr, Jan N. van Rijn, Haozhe Sun, Joaquin Vanschoren, and Phan Anh Vu. Meta-
 768 album: Multi-domain meta-dataset for few-shot image classification. In Sanmi Koyejo,
 769 S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in*
 770 *Neural Information Processing Systems 35: Annual Conference on Neural Information*
 771 *Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - Decem-*
 772 *ber 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/1585da86b5a3c4fb15520a2b3682051f-Abstract-Datasets_and_Benchmarks.html.

773 Cyrille Violle, Marie-Laure Navas, Denis Vile, Elena Kazakou, Claire Fortunel, Irène Hum-
 774 mel, and Eric Garnier. Let the concept of trait be functional! *Oikos*, 116(5):882–892,
 775 2007. URL <https://nsojournals.onlinelibrary.wiley.com/doi/10.1111/j.0030-1299.2007.15559.x>.

776 Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd
 777 birds-200-2011 dataset. 2011.

778 Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing
 779 Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men,
 780 Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-
 781 language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*,
 782 2024. URL <https://arxiv.org/abs/2409.12191>.

783 Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Axbench: Steering LLMs? even simple baselines
 784 outperform sparse autoencoders. In *Forty-second International Conference on Machine Learning*,
 785 2025. URL <https://openreview.net/forum?id=K2CckZjNy0>.

786 Zeyu Yun, Yubei Chen, Bruno A Olshausen, and Yann LeCun. Transformer visualization via
 787 dictionary learning: contextualized embedding as a linear superposition of transformer factors.
 788 *arXiv preprint arXiv:2103.15949*, 2021. URL <https://arxiv.org/abs/2103.15949>.

789

790

791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

810 APPENDICES
811

812 This supplementary material provides additional details omitted in the main text.
813

814 **Contents**
815

816	Appendices	16
817		
818	A Limitations	17
819		
820	B Comprehensive Results	17
821		
822	C System Prompts	18
823		
824	D Experimental Setup	20
825	D.1 Hyperparameter Configuration	20
826	D.2 Downstream Evaluation	20
827	E Crowdsourcing Details	20
828		
829	F Dataset Examples	21
830		
831	G Additional Neuron Activation Analysis	21
832		
833	H Additional Dataset Ablation Examples	25
834	H.1 MLLM + SAE vs. MLLM-only baseline	25
835	H.2 Qwen-2.5-VL-7B vs. Qwen-2.5-VL-72B	27
836	H.3 Multiple vs. Single Image per Latent	27
837	I Feature Detector Ablations	27
838		
839	J LLM Usage Details	29
840		
841	K Ecology Applications	29
842		
843		
844		
845		
846		
847		
848		
849		
850		
851		
852		
853		
854		
855		
856		
857		
858		
859		
860		
861		
862		
863		

864 A LIMITATIONS

866 We assume that the dense features from the backbone image foundation model encode morphology-
 867 relevant signals. If these representations are biased toward generic visual concepts, important
 868 biological traits may be underrepresented. The SAE discovers latent factors that are spatially and
 869 semantically coherent, but some latents might correspond to multiple co-occurring traits (e.g., “eloni-
 870 gated + thin”). This can make it difficult to disentangle fine-grained trait attributes or compositional
 871 traits. Trait descriptions generated with smaller MLLMs like Qwen-2.5-VL-7B are susceptible to
 872 hallucination, particularly when prompted with noisy or background-dominated patches. Also, eval-
 873 uating trait correctness at scale remains a challenge due to the absence of ground-truth morphological
 874 trait annotations.

875 Recent work (Kantamneni et al., 2025; Wu et al., 2025) has highlighted the limitations of SAEs,
 876 showing that they do not consistently outperform simpler baselines on downstream tasks. However,
 877 we do not use SAEs for steering or sparse probing in LLMs, but rather as a pragmatic tool for
 878 proposing spatially localized, candidate part detectors in DINOv2 features that can be grounded
 879 to image patches and then described by an MLLM. We mitigate some known SAE limitations by
 880 (i) applying species-contrastive ranking and frequency thresholds to filter out spurious latents, (ii)
 881 enforcing multi-image consistency (traits must recur across many instances of the same species), and
 882 (iii) evaluating the resulting traits both via expert ratings and via downstream transfer to in-the-wild
 883 Insects classification. In other words, we do not assume that SAE features are the true underlying
 884 traits; instead, we treat them as a useful decomposition that is subsequently empirically validated and
 885 filtered.

886 B COMPREHENSIVE RESULTS

888 The comprehensive results with standard deviation for ratings for various ablations are given in
 889 Table B.1, Table B.2, and Table B.3, respectively.

891 Table B.1: Incorporating latent-specific patches significantly improves the quality of trait descriptions.
 892 Including multiple images in the prompt encourages MLLMs to focus on the traits common across all
 893 images, at the cost of more tokens per query. Using multiple images with SAE-extracted bounding
 894 boxes leads to improved precision, as better ratings indicate. The experimental setup uses Qwen2.5-
 895 VL-72B as MLLM, a normalized frequency threshold of (t_{freq}) = 3e-3, and 1,000 input images.

Method	# Images	# Tokens	# Images	# Traits	Avg. Rating
MLLM	1	413	—	—	3.00 (± 0.71)
MLLM	3	940	—	—	3.15 (± 0.54)
MLLM + SAE	1	411	460	9,435	3.84 (± 0.63)
MLLM + SAE	3	1,072	460	7,897	3.91 (± 0.92)

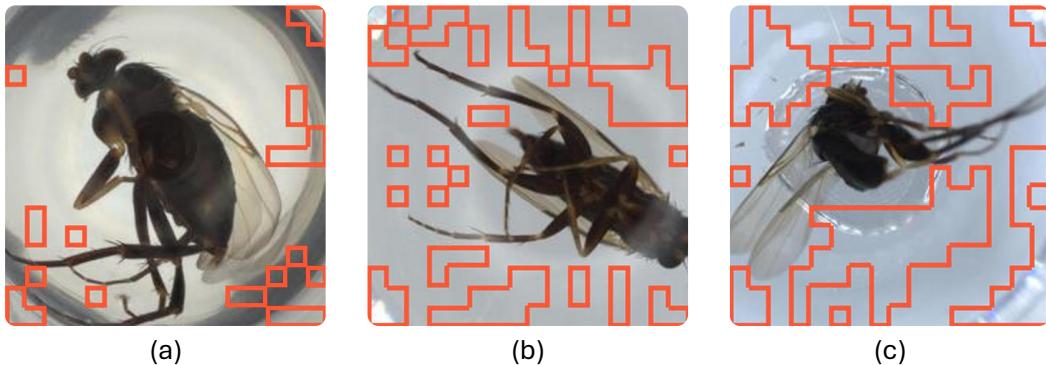
903 Table B.2: SAEs often trade off between reconstruction error (MSE) and sparsity (L_0). We investigate
 904 the effect of choosing between different balances of these errors. We find that lower sparsity performs
 905 better for both values of frequency threshold (t_{freq}). A lower value of the sparsity coefficient (α) leads
 906 to lower MSE and thus better reconstruction. It improves the coverage of latents, leading to better
 907 recall. The experimental setup uses an input dataset of 1,000 images.

Method	α	t_{freq}	SAE MSE	SAE L_0	# Images	# Traits	Avg. Rating
MLLM+SAE	2e-4	1e-2	8.8e-3	1,081.1	60	60	3.84 (± 0.70)
MLLM+SAE	4e-4	3e-3	2.7e-2	690.4	460	7,897	3.91 (± 0.92)
MLLM+SAE	4e-4	1e-2	2.7e-2	690.4	20	20	3.58 (± 1.05)
MLLM+SAE	8e-4	3e-3	5.4e-2	242.2	458	3,060	3.87 (± 0.83)

916 **MLLM Quality Ablations.** To evaluate the impact of model scale on morphological trait generation,
 917 we compare descriptions produced by Qwen2.5-VL-7B and Qwen2.5-VL-72B (Wang et al., 2024)

Table B.3: We investigate the effect of the verbalizer **MLLM** for morphological trait extraction for both the MLLM-only and MLLM + SAE models. We observe that **GPT-5-mini** achieves the highest average rating, outperforming both open Qwen-2.5-VL variants by a substantial margin. The larger Qwen-2.5-VL-72B model (Wang et al., 2024) consistently obtains better ratings than its 7B counterpart. We note that GPT-5 mini and Qwen-2.5-VL-7B models might lead to false positives due to hallucination while extracting common traits in three input SAE-annotated images (Figure B.1). In contrast, the Qwen2.5-VL-72B model demonstrates improved robustness, avoiding such hallucinations and yielding more accurate trait descriptions. The experimental setup uses an input dataset of 20K images and $t_{\text{freq}} = 1e-2$.

Method	MLLM	# Images	# Traits	Avg. Rating
MLLM	Qwen-2.5 VL 7B	476	536	2.85 (± 0.67)
MLLM	Qwen-2.5 VL 72B	371	411	3.15 (± 0.54)
MLLM + SAE	Qwen-2.5 VL 7B	478	538	2.90 (± 1.39)
MLLM + SAE	Qwen-2.5 VL 72B	358	370	3.58 (± 1.05)
MLLM + SAE	GPT-5 mini	478	538	4.04 (± 0.45)



Qwen2.5-VL-7B: Wing: Transparent, delicate, slightly curved, light brownish color, fine texture.

GPT-5 mini: - Wing: translucent, pale, elongated, Legs: dark, slender, jointed, - Body

(thorax/abdomen): dark brown, rounded, glossy.

Qwen2.5-VL-72B: None: No shared visible traits across all three regions.

Figure B.1: Comparison of morphological trait description quality between Qwen2.5-VL-7B, GPT-5 mini, and Qwen2.5-VL-72B for *Diplonevra nitidula*. Each red box highlights a region selected by SAE neurons with high activation, indicating regions used for prompting the MLLM + SAE. The Qwen2.5-VL-72B model correctly recognizes the background context and refrains from hallucinating visible traits, suggesting improved spatial grounding.

when prompted with latent-indexed image patches (Table B.3). The larger 72B model consistently receives higher human evaluation scores than its 7B counterpart. In one illustrative example, Qwen2.5-VL-72B correctly identifies a red-boxed region as background, while the 7B model incorrectly hallucinates a body part description (Figure B.1). These results suggest that larger models exhibit improved spatial grounding and are more reliable in avoiding false positive trait attributions.

C SYSTEM PROMPTS

The prompts used for the MLLM + SAE model are shown in Figure C.2 and Figure C.3, corresponding to the multi-image and single-image settings, respectively. For comparison, the prompts for the MLLM-only baseline are provided in Figure C.4 (multi-image) and Figure C.5 (single-image).

972
 973
 974
 975 You are given three images of insects, each with multiple red bounding boxes highlighting specific
 976 regions.
 977
 978 For each image:
 979 1. For every highlighted region, determine whether it contains a visible insect body part or just
 980 background. If it is mostly background, respond with "background".
 981
 982 2. If it contains a visible body part, identify which part it is (e.g., leg, wing, antenna), and describe
 983 its visible morphological traits: shape, size, color, texture, and any distinct markings. Use only the
 984 visual information present in the image.
 985 After analyzing all three highlighted regions in images:
 986
 987 3. Identify and list the morphological traits that are **common across all three regions, solely**
 988 **based on what is visible in all images.**
 989
Important Instructions:
 990 - Do not infer or assume information that is not directly observable. Avoid adding external
 991 knowledge.
 992 - Use only what is clearly visible.
 993 - Be concise. Limit the total response to under 200 tokens.
 994
Output Format:
 995 - [Image 1]:
 996 - [Body Part]: [Visible trait]
 997 - [Image 2]:
 998 - [Body Part]: [Visible trait]
 999 - [Image 3]:
 1000 - [Body Part]: [Visible trait]
 1001 - [Common Traits Across All Three Images]:
 1002 - [Body Part]: [Shared visible trait]
 1003
 1004

Figure C.2: Prompt for MLLM + SAE (multiple images)

1005
 1006
 1007
 1008
 1009
 1010 You are given an image of an insect with multiple red bounding boxes overlaid on it, highlighting a
 1011 specific region.
 1012
 1013 1. Determine whether the highlighted region contains a visible body part of the insect or only the
 1014 background. If it appears to be background, respond with "background".
 1015
 1016 2. If it contains a visible body part, identify which part it is. Then, briefly describe the observable
 1017 morphological features - such as shape, size, color, texture, or distinct markings - **based solely on**
 1018 **what is visible in the image.**
 1019
IMPORTANT: Do not infer or assume information that is not directly observable. Avoid adding
 1020 external knowledge.
 1021
 1022

Figure C.3: Prompt for MLLM + SAE (single image)

1026
 1027 You are given three images of insects. Your task is to visually analyze them and extract observable
 1028 morphological traits.
 1029
 1030 1. Identify the visible body parts of the insect (e.g., head, thorax, abdomen, legs, wings, antennae),
common in all three images.
 1031 2. For each part, identify its morphological features - such as shape, size, color, texture, or distinct
 1032 markings.
 1033 3. After analyzing all three images individually, list the morphological traits that are **common across**
 1034 **all three insects. Only output traits that are visibly consistent across all images.**
 1035
IMPORTANT: Do not infer or assume information that is not directly observable. Avoid adding external
 1036 knowledge.
 1037

Figure C.4: Prompt for MLLM-only baseline (multiple images)

1038
 1039
 1040
 1041 You are given an image of an insect specimen. Your task is to visually examine the insect and describe
 1042 its observable morphological traits.
 1043
 1044 1. Identify the visible body parts of the insect (e.g., head, thorax, abdomen, legs, wings, antennae).
 1045 2. For each part, briefly describe the observable morphological features - such as shape, size, color,
 1046 texture, or distinct markings - **based solely on what is visible in the image.**
 1047
IMPORTANT:
 1048 1. Do not infer or assume information that is not directly observable. Avoid adding external knowledge.
 1049 2. Keep your response concise and under 200 tokens.
 1050

Figure C.5: Prompt for MLLM-only baseline (single image)

D EXPERIMENTAL SETUP

D.1 HYPERPARAMETER CONFIGURATION

1054
 1055 Table D.4 summarizes all hyperparameters used for SAE training and dataset generation. We
 1056 experiment with different learning rate values and choose $1e-3$ based on qualitative inspection of
 1057 learned traits. All experiments were conducted on NVIDIA H100 GPUs. SAE training required
1058 approximately 11 hours, while the dataset generation took 183 hours using a single-process setup.

D.2 DOWNSTREAM EVALUATION

1060
 1061 For downstream evaluation, we use the Insects dataset Ullah et al. (2022), which consists of volunteer
 1062 field photos of live insects interacting with flowers and foliage, often partially occluded, in diverse
 1063 poses, backgrounds, and viewing distances. This introduces multiple distribution shifts (background
 1064 clutter, illumination, pose, occlusion, and scale) beyond the lab setting. We fine-tune BioCLIP in a
 1065 standard image–text contrastive manner, where the text input is a caption that concatenates the species
 1066 name with the trait description. Concretely, we use prompts of the form “A photo of <species-name>
 1067 with <trait-description>.”

E CROWDSOURCING DETAILS

1068
 1069 All trait description ratings were performed solely by the authors of this paper, who voluntarily
 1070 participated in the evaluation. The IRB indicated that our research is exempt and does not require
 1071 approval. The evaluation rubric is shown in Table E.6.

1080

Table D.4: Hyperparameters for SAE training and filtering.

1081

1082

1083

1084

1085

1086

1087

1088

1089

1090

1091

1092

1093

Hyperparameter	Value
Hidden Width	24,576 ($32 \times$ expansion)
Sparsity Coefficient α	{2e-4, 4e-4, 8e-4}
Sparsity Coefficient Warmup	500 steps
Batch Size	16,384
Learning Rate η	{5e-4, 1e-3}
Learning Rate Warmup	500 steps
Activation threshold $t_{\text{activation}}$	0.9
ViT layer ID	10

1094

1095

1096

1097

1098

1099

1100

1101

1102

Table D.5: Dataset statistics. On average, each image is associated with 4.2 trait samples.

1103

1104

Metric	Value
# Species	736
# Genera	417
# Unique images	19.1K
# Samples	80.8K

1105

1106

1107

1108

F DATASET EXAMPLES

1109

1110

1111

1112

1113

1114

1115

1116

1117

1118

1119

1120



1121

1122

1123

Figure F.6: Example 1 from BIOSCAN-TRAITS: “- Wing: Transparent, elongated, with visible veins.
- Antenna: Thin, segmented, light brown”.

1124

1125

1126

1127

1128

1129

G ADDITIONAL NEURON ACTIVATION ANALYSIS

1130

1131

1132

1133

Similar to Section 4.3, we analyze additional cases of the top-activating neurons (or latent dimensions) in the SAE to investigate whether they correspond to meaningful morphological traits (Figure G.11). For instance, we observe that within the SAE, neuron 4040 consistently activates on the thorax, while neuron 16584 responds to the leg-body junction, highlighting spatially grounded morphological regions.

1134

1135

1136

1137

1138

1139

1140

1141

1142

1143

1144

1145

1146

1147

1148

1149

1150

1151

1152

1153

1154

1155

1156

1157

1158

1159

1160

1161

1162

1163

1164

1165

1166

1167

1168

1169

1170

1171

1172

1173

1174

1175

1176

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187

Table E.6: Example-based rubric for evaluating trait descriptions.

Score	Example Image	Evaluation Criteria
5		<p>Completely Correct — Body part correctly identified; all traits visibly match (color, texture, shape, size); no hallucinations.</p> <p>Example: “[Leg]: Thin, elongated, light brown, segmented.”</p>
4		<p>Mostly Correct — Body part is correct; Most traits are accurate, one minor imprecision.</p> <p>Example: “Leg: Thin, segmented, translucent; jointed structure” (leg is not translucent, but other traits are correct)</p>
3		<p>Partially Correct — Body part correct; 1–2 traits vague or incorrect.</p> <p>Example: “[Leg]: Thick, black, elongated.” (body part is correct, but leg is thin and brown.)</p>
2		<p>Mostly Incorrect — Incorrect body part or major trait mismatches.</p> <p>Example: “[Segmented]: All parts are visibly segmented” (the body part is missing, the segmented trait is correct though)</p>
1		<p>Completely Incorrect — Hallucinated or wrong body part.</p> <p>Example: “[Antennae]: dark brown.” (No antennae are visible)</p>

1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198



1199 Figure F.7: Example 2 from BIOSCAN-TRAITS: “- [Leg]: Thin, elongated, light brown, segmented”.
1200

1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211



1212 Figure F.8: Example 3 from BIOSCAN-TRAITS: “- Wing: Transparent, elongated, with visible veins.
1213 - Antenna: Thin, segmented, dark brown”.
1214

1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225



1226 Figure F.9: Example 4 from BIOSCAN-TRAITS: “- Wing: Brown, translucent, folded, with visible
1227 veins”.
1228

1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239



1240 Figure F.10: Example 5 from BIOSCAN-TRAITS: “- Antenna: Thin, elongated, segmented, dark
1241 color”.
1242

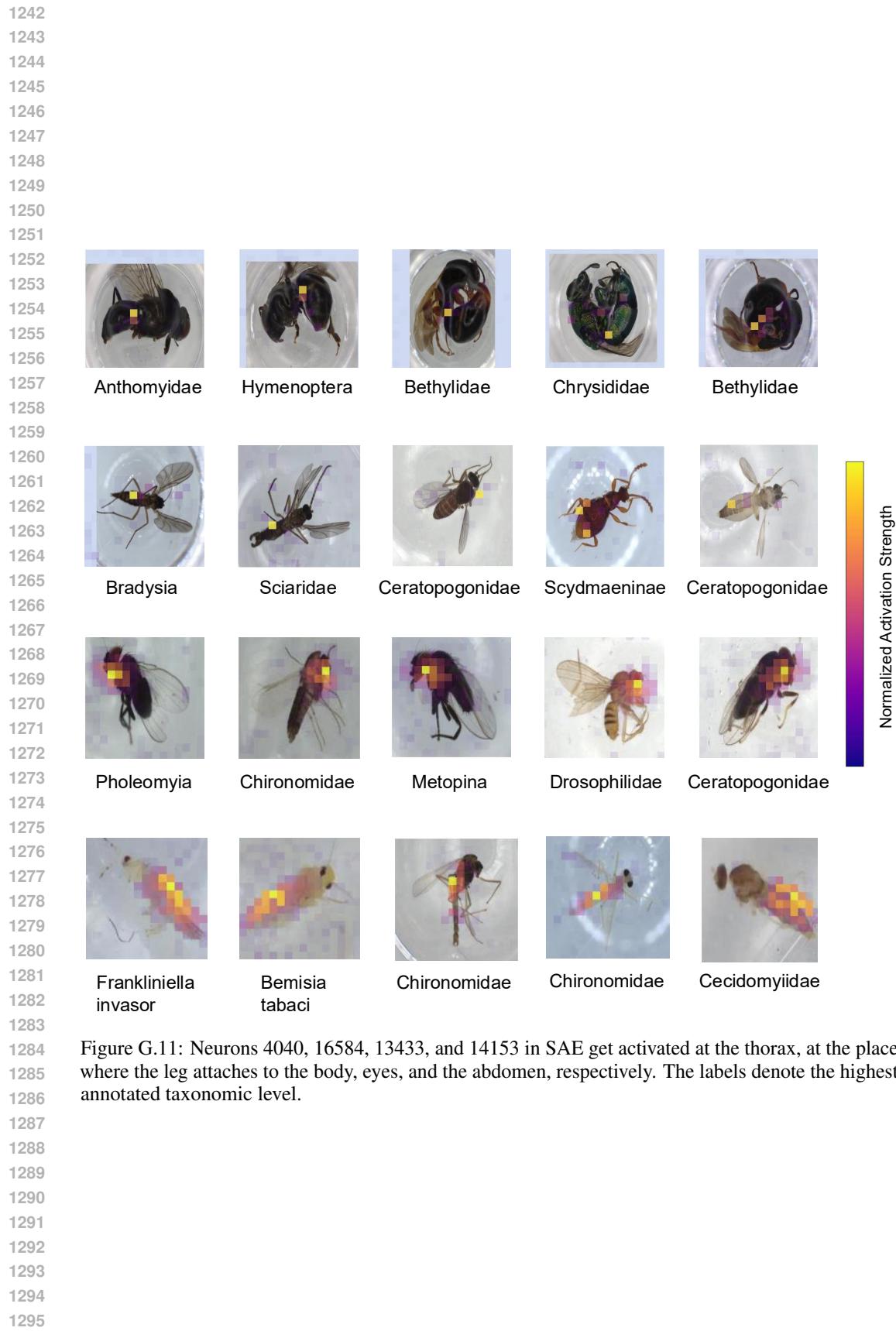
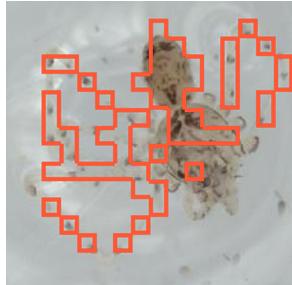


Figure G.11: Neurons 4040, 16584, 13433, and 14153 in SAE get activated at the thorax, at the place where the leg attaches to the body, eyes, and the abdomen, respectively. The labels denote the highest annotated taxonomic level.

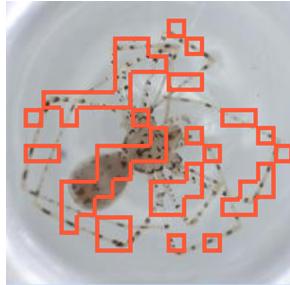
1296 **H ADDITIONAL DATASET ABLATION EXAMPLES**
 1297
 1298
 1299

1300 **H.1 MLLM + SAE vs. MLLM-ONLY BASELINE**
 1301
 1302

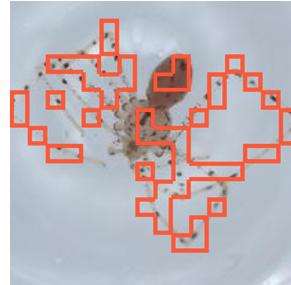
Figure H.12- H.15 present additional examples comparing the salient morphological trait descriptions generated by the MLLM-only baseline versus MLLM + SAE.



(a)

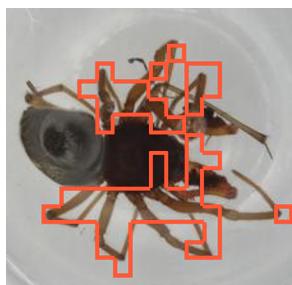


(b)



(c)

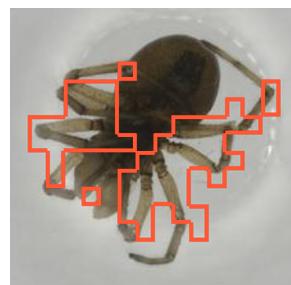
Figure H.12: Comparison of salient morphological trait description generation using a just MLLM vs. MLLM + SAE for *Scytodes intricata*. Each red box highlights a region selected by SAE neurons with high activation, indicating regions used for prompting the MLLM + SAE. The use of SAE helps MLLMs focus on salient morphological traits rather than general descriptions of all body parts. **MLLM: Common Morphological Traits Across All Three Images:** 1. *Head*: - *Shape*: Rounded, slightly elongated. - *Color/Texture*: Light-colored with darker speckled patterns. - *Distinct Markings*: Dark spots or patches scattered across the surface. 2. *Thorax*: - *Shape*: Small and compact, merging seamlessly with the head. - *Color/Texture*: Similar light coloration as the head, with dark speckles. - *Distinct Markings*: Speckled pattern continues from the head. 3. *Abdomen*: - *Shape*: Oval-shaped, larger than the head and thorax combined. - *Color/Texture*: Light base color with a dense covering of dark speckles. - *Distinct Markings*: Dark speckles form a mottled appearance. **MLLM + SAE:** [Thin, elongated structure]: Segmented appearance, brownish/dark coloration.



(a)

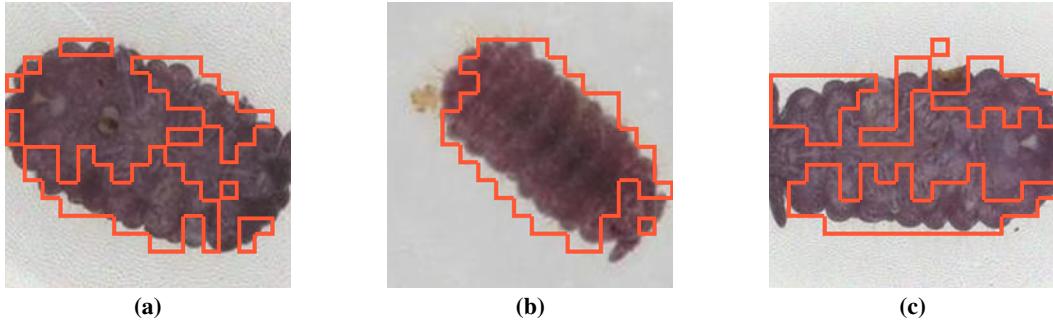


(b)



(c)

Figure H.13: Comparison of salient morphological trait description generation using a just MLLM vs. MLLM + SAE for *Erigone psychrophila*. Each red box highlights a region selected by SAE neurons with high activation, indicating regions used for prompting the MLLM + SAE. The use of SAE helps MLLMs focus on salient morphological traits rather than general descriptions of all body parts. **MLLM: Common Morphological Traits Across All Three Images:** 1. *Head* - *Shape*: Rounded and slightly flattened. - *Size*: Small relative to the body. - *Color*: Light brown or beige. - *Texture*: Appears smooth with no distinct markings. 2. *Thorax* - *Shape*: Slightly elongated and oval. - *Size*: Medium-sized, connecting the head and abdomen. - *Color*: Light brown or beige, similar to the head. - *Texture*: Smooth, no distinct markings. 3. *Abdomen* - *Shape*: Large, rounded, and bulbous. - *Size*: Significantly larger than the head and thorax. - *Color*: Darker brown compared to the head and thorax. - *Texture*: Smooth, with a glossy appearance. **MLLM + SAE:** [Leg]: Thin, elongated, brown, segmented.

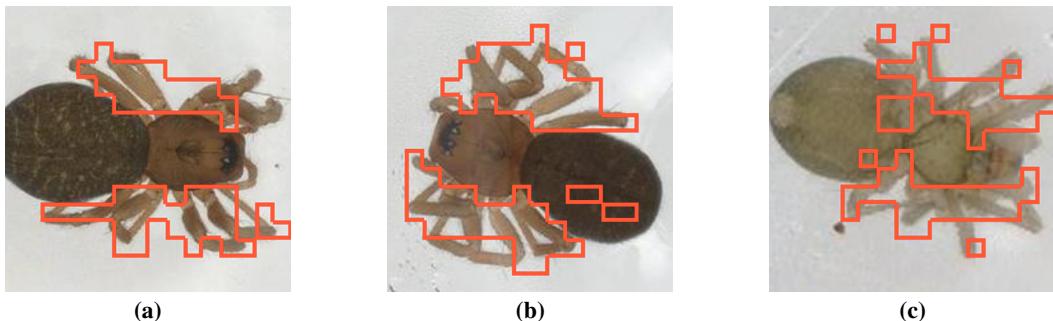


1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379

Figure H.14: Comparison of salient morphological trait description generation using a just MLLM vs. MLLM + SAE for *Morulina thulensis*. Each red box highlights a region selected by SAE neurons with high activation, indicating regions used for prompting the MLLM + SAE. The use of SAE helps MLLMs focus on salient morphological traits rather than general descriptions of all body parts.

MLLM: Common Morphological Traits Across All Three Images: 1. *Body Segmentation*: - *Shape*: The body is segmented into multiple rounded sections. - *Size*: The segments appear uniform in size relative to one another. - *Color*: The segments are dark purple or maroon. - *Texture*: The surface appears rough with small protrusions or bumps on each segment. 2. *Head*: - *Shape*: The head is slightly smaller than the other segments and is located at one end of the body. - *Color*: The head matches the dark purple or maroon color of the rest of the body. - *Distinct Markings*: A lighter-colored area is visible near the front of the head, possibly indicating sensory organs or mouthparts. 3. *Antennae*: - *Shape*: Two short, thin antennae are visible extending from the head.

MLLM + SAE: Body: Dark brown, textured surface; irregular shape.



1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403

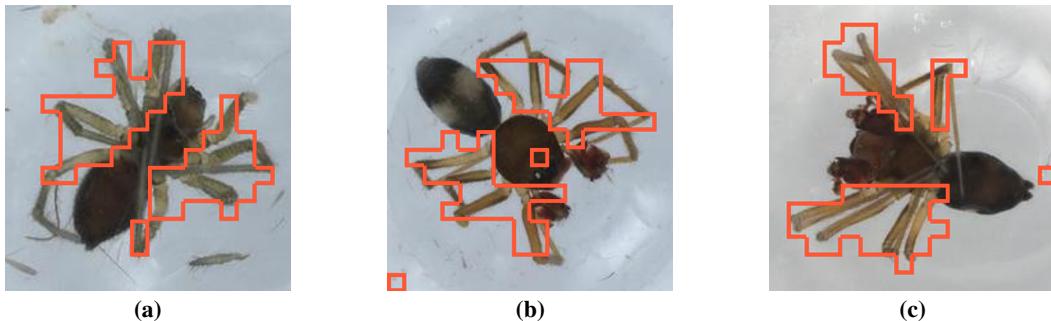
Figure H.15: Comparison of salient morphological trait description generation using a just MLLM vs. MLLM + SAE for *Islandiana cristata*. Each red box highlights a region selected by SAE neurons with high activation, indicating regions used for prompting the MLLM + SAE. The use of SAE helps MLLMs focus on salient morphological traits rather than general descriptions of all body parts.

MLLM: Common Morphological Traits Across All Three Images: 1. *Head - Shape*: Rounded. - *Size*: Small relative to the body. - *Color*: Light brown or beige. - *Texture*: Appears smooth with some fine hairs. 2. *Thorax - Shape*: Not distinctly visible as a separate segment; appears fused with the head. - *Size*: Small, integrated with the head region. - *Color*: Similar to the head, light brown or beige. - *Texture*: Smooth with fine hairs. 3. *Abdomen - Shape*: Oval and rounded. - *Size*: Larger than the head-thorax region. - *Color*: Darker compared to the head-thorax, ranging from dark brown to black. - *Texture*: Smooth with possible fine hairs.

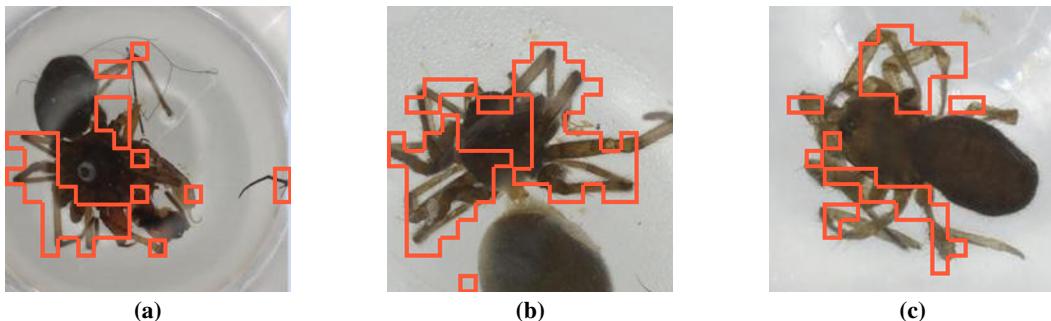
MLLM + SAE: [Leg]: Thin, elongated, brown, segmented.

1404 H.2 QWEN-2.5-VL-7B vs. QWEN-2.5-VL-72B
 1405

1406 Figure H.16- H.17 present additional examples comparing the salient morphological trait descriptions
 1407 generated by Qwen-2.5-VL-7B vs. Qwen-2.5-VL-72B as the backbone MLLM for MLLM + SAE.
 1408 The larger Qwen2.5-VL-72B model accurately identifies the insect's body parts and avoids the
 1409 hallucinations observed in its 7B counterpart.



1421 Figure H.16: Comparison of morphological trait description quality between Qwen2.5-VL-7B and
 1422 Qwen2.5-VL-72B for *Agyneta straminicola*. Each red box highlights a region selected by SAE
 1423 neurons with high activation, indicating regions used for prompting the MLLM + SAE. **Qwen2.5-**
 1424 **VL-7B:** [Body Part]: Wing - Shared visible trait: Large, translucent, with a slightly darker margin,
 1425 smooth texture. **Qwen2.5-VL-72B:** [Leg]: Thin, elongated, segmented, light brown. The larger
 1426 model correctly identifies the highlighted body part of the insect.



1427
 1428 Figure H.17: Comparison of morphological trait description quality between Qwen2.5-VL-7B and
 1429 Qwen2.5-VL-72B for *Erigone psychrophila*. Each red box highlights a region selected by SAE
 1430 neurons with high activation, indicating regions used for prompting the MLLM + SAE. **Qwen2.5-**
 1431 **VL-7B:** Wing - [Shared visible trait]: Translucent, with a darker margin. **Qwen2.5-VL-72B:** [Leg]:
 1432 Thin, elongated, dark brown, segmented. - [Antenna]: Thin, long, dark brown, smooth. The larger
 1433 model correctly identifies the highlighted body part of the insect.

1434 H.3 MULTIPLE VS. SINGLE IMAGE PER LATENT
 1435

1436 Figure H.18-H.19 present additional examples comparing the salient morphological trait descriptions
 1437 generated using a single image versus multiple images for MLLM + SAE. This consensus-driven
 1438 trait extraction encourages the model to focus on consistent traits and leads to improved precision.

1439 I FEATURE DETECTOR ABLATIONS
 1440

1441 We use DINOv2-base (ViT-B/14) (Oquab et al., 2024) as our feature extractor, motivated by prior
 1442 work showing its effectiveness in producing high-quality SAE representations (Stevens et al., 2025;
 1443 Pach et al., 2025). To validate this choice, we conducted preliminary experiments on a 1000-species
 1444 benchmark derived from BIOSCAN-5M (20 train / 30 test images per species), comparing CLIP
 1445 ViT-B/16 (Caron et al., 2021) and DINOv2-base features (Table I.7).

1458

1459

1460

1461

1462

1463

1464

1465

1466

1467

1468

1469

1470



Figure H.18: Comparison of salient morphological trait description generation using a single image vs. three images for *Deltocephalus fuscinervosus*. Each red box highlights a region selected by SAE neurons with high activation, indicating regions used for prompting the MLLM + SAE. Using multiple images with SAE-extracted bounding boxes leads to dramatically improved precision. **(a)** **Single image:** The highlighted area appears to encompass the thorax and part of the abdomen of the insect. **Observable Morphological Features:** - **Shape:** The thorax is somewhat rectangular and connects to the abdomen, which tapers slightly towards the end. - **Size:** The thorax is relatively large compared to the head and legs, while the abdomen is elongated. - **Color:** The thorax and abdomen have a dark brown coloration with lighter, possibly yellowish or beige, patterns or markings. - **Texture:** The surface appears to have fine hairs or setae, giving it a slightly textured look. - **Distinct Markings:** There are lighter, irregularly shaped markings on the darker background of the thorax and abdomen. **(a, b, c) Multiple images:** Fine Hairs: Present on legs and body, small and sparse.

1482

1483

1484

1485

1486

1487

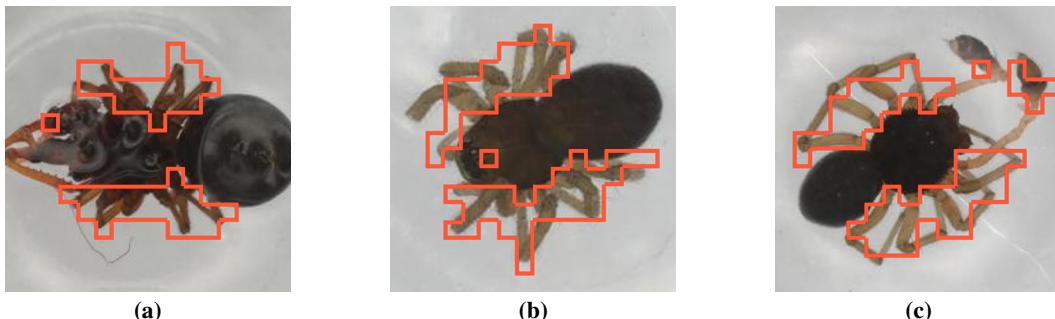


Figure H.19: Comparison of salient morphological trait description generation using a single image vs. three images for *Erigone arctophylacis*. Each red box highlights a region selected by SAE neurons with high activation, indicating regions used for prompting the MLLM + SAE. Using multiple images with SAE-extracted bounding boxes leads to dramatically improved precision. **(a)** **Single image:** The highlighted area encompasses the legs and part of the thorax of the insect. **Observable Morphological Features:** - **Shape:** The legs are elongated and segmented, with joints visible at various points along their length. The thorax is partially visible and appears rounded. - **Size:** The legs are relatively thin compared to the overall size of the insect. The thorax section within the highlighted area is broader than the legs. - **Color:** The legs and thorax appear to be a dark brown or black color. - **Texture:** The surface of the legs seems somewhat smooth but with fine hairs or setae visible upon close inspection. - **Distinct Markings:** There are no distinct markings or patterns visible on the legs or thorax within the highlighted area. **(a, b, c) Multiple images:** [Leg]: Elongated, brown, segmented.

1510

1511

1512 We observed that DINOv2-base substantially outperforms CLIP ViT-B/16, using the kNN classifier.
 1513 Based on these results, we selected DINOv2-base as our backbone. Following prior work (Stevens
 1514 et al., 2025), we extract features from the penultimate layer of the ViT for SAE training.
 1515

1516 Table I.7: Species classification accuracy on 1000-species benchmark derived from BIOSCAN-5M
 1517 (20 train / 30 test images per species). DINOv2-base substantially outperforms CLIP ViT-B/16, using
 1518 the kNN classifier.

Model	Top1 Accuracy (%)
CLIP ViT-B/16, kNN	24.57
SigCLIP ViT-B/16, kNN	29.68
DINOv2-base, kNN	41.28

J LLM USAGE DETAILS

We utilized large language models (LLMs) to aid in the writing and editing of this paper. Their role within our trait-generation pipeline, specifically the use of multimodal LLMs (MLLMs), is described in Section 3.

K ECOLOGY APPLICATIONS

Below, we outline several concrete ways in which ecologists can leverage the proposed trait-generation pipeline:

- **Expanding trait databases:** Building trait databases by hand using domain experts is time-consuming. An automated tool can quickly add thousands of traits from existing images, populating databases or filling gaps. This helps ecologists who rely on traits (for example, to model species’ niches or ecosystem roles) by providing many more data points.
- **Enabling new analyses:** With rich trait labels attached to images, researchers can study correlations between morphology and environment or behavior at scale. For instance, you could analyze how wing shapes vary across climates, or link body color patterns to predation risk. Traits explain ecological patterns better than just species names, and an automated pipeline makes these analyses feasible on large collections.
- **Boosting identification tools:** As shown with BioCLIP, trait-annotated images can improve automatic species-identification models. Models trained on trait captions learn more nuanced visual cues, making them more robust to new specimens or image conditions.

Overall, our pipeline provides a scalable way to inject expert-like knowledge (descriptions of body parts) into machine learning without manual annotation . By turning images into meaningful trait statements, it bridges the gap between digitized specimens and quantitative trait databases, supporting a wide range of biodiversity and ecological research.

1552
 1553
 1554
 1555
 1556
 1557
 1558
 1559
 1560
 1561
 1562
 1563
 1564
 1565