Mind the (Data) Gap: Evaluating Vision Systems in Small Data Applications

Anonymous Author(s)
Affiliation
Address
email

Abstract

The practical application of AI tools for specific computer vision tasks relies on the "small-data regime" of hundreds to thousands of labeled samples. This small-data regime is vital for applications requiring expensive expert annotations, such as ecological monitoring, medical diagnostics or industrial quality control. We find, however, that computer vision research has ignored the small data regime as evaluations increasingly focus on zero- and few-shot learning. We use the Natural World Tasks (NeWT) benchmark to compare multi-modal large language models (MLLMs) and vision-only methods across varying training set sizes. MLLMs exhibit early performance plateaus, while vision-only methods improve throughout the small-data regime, with performance gaps widening beyond 10 training examples. We provide the first comprehensive comparison between these approaches in small-data contexts and advocate for explicit small-data evaluations in AI research to better bridge theoretical advances with practical deployments.

1 Introduction

AI research has increasingly favored evaluating new methods primarily through zero-shot and few-shot benchmarks [7, 9, 19, 32, 46]. This evaluation approach is driven by the compelling promise of strong generalization with minimal examples. However, this focus on zero—and few-shot learning neglects a pervasive and essential scenario: the *small-data regime*, characterized by datasets containing roughly dozens to a few thousand labeled samples (see Fig. 1a). This regime is critical for numerous real-world applications where extensive labeled data collection remains costly and challenging, such as ecological monitoring [5, 41], medical diagnostics [14], and industrial quality control [43]. Our community's decreased attention to rigorous small-data evaluations is a significant oversight. By optimizing primarily for zero-shot and few-shot performance, we risk developing methods ill-suited for practical scenarios where moderate data availability is typical. To address this gap, evaluations of new methods should explicitly include small-data assessments.

To systematically evaluate the small-data regime, we use the Natural World Tasks [NeWT; 42] benchmark, which is specifically designed for challenging fine-grained ecological classification tasks requiring expert annotation. Using this benchmark, we compare multimodal large language models (MLLMs) and vison-only methods across varying training set sizes. Using ecological classification tasks as representative test cases, we analyze model performance and scaling behavior. Our findings highlight significant limitations of current MLLMs, notably early performance plateaus, in contrast to sustained performance improvements observed in vision-only methods as dataset sizes increase within the small-data regime (see Fig. 1b). While our study utilizes ecological tasks as a convenient testbed due to the availability of the NeWT benchmark in the small-data regime, our argued position extends beyond ecology to the broader field of computer vision applications where limited labeled data is a common constraint. In this work, we:

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

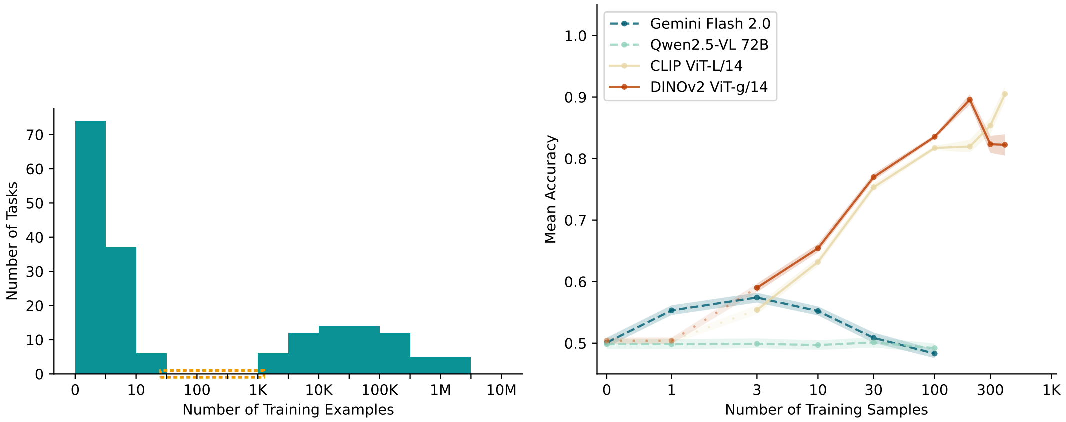

Figure 1: **Left:** Unique evaluation tasks used in recent language and vision research [1, 2, 4, 10, 16, 18, 25, 27, 28, 30, 32, 40, 47] summarized by the number of training samples per task. Note how few evaluations use between 10 and 1,000 labeled training samples. We collect this data manually. **Right:** Mean NeWT task performance as a function of number of labeled examples for multimodal large language models (MLLMs) and vision-only models combined with support vector machines (SVMs). MLLMs leverage labeled examples by including additional labeled examples in the prompt (few-shot prompting). Vision models leverage training examples by fitting an SVM to frozen image embeddings. Vision models with SVMs improve with additional training data and consistently outperform MLLMs with 10 or more labeled samples. Note the log scale for training data. Shaded areas indicate bootstrapped 95% confidence intervals.

1. Emphasize the critical-but-neglected small-data regime and advocate for its inclusion in AI research benchmarks.
2. Conduct the first comparison of foundation models versus vision-only methods within the small-data regime.

This work profiles performance patterns across model types and data scales. While our findings can inform new research directions, we avoid model selection advice as optimal approaches depend on application-specific constraints. We aim to present empirical evidence highlighting the need for small-data evaluations in AI research.

2 Background & Related Work

We highlight a gap in evaluation practices, discuss trends and visually summarize the *small-data gap* in Fig. 1a.

Evaluation Trends in AI Research. Recent computer vision methods [30, 32], and (multimodal) large language models [(M)LLMs; 2, 3, 11, 29] primarily evaluate performance using zero-shot or few-shot benchmarks [19, 38, 46]. These benchmarks emphasize generalization with extremely limited examples, reflecting a trend toward model robustness with minimal fine-tuning.

Evaluation Trends in Ecological Computer Vision. While ecological computer vision has begun adopting multimodal and foundation models for tasks such as species identification [41, 42], the evaluations still frequently rely on fixed data splits or zero/few-shot scenarios. For instance, ecological benchmarks such as iNat2021 [42] or iWildCam [6, 23] evaluate systems on 10K+ labeled examples without systematically exploring performance scaling within moderate-sized training sets. Recent specialized foundation models [17, 33, 34] demonstrate interest in domain-specific representations, yet small-data evaluations remain uncommon.

Small-Data Gap Despite the practical importance of evaluating methods with tens to thousands of labeled examples, a regime typical in real-world ecological, medical, and industrial applications [8, 24], current methods research neglect systematic evaluation at these scales (see Fig. 1a). This

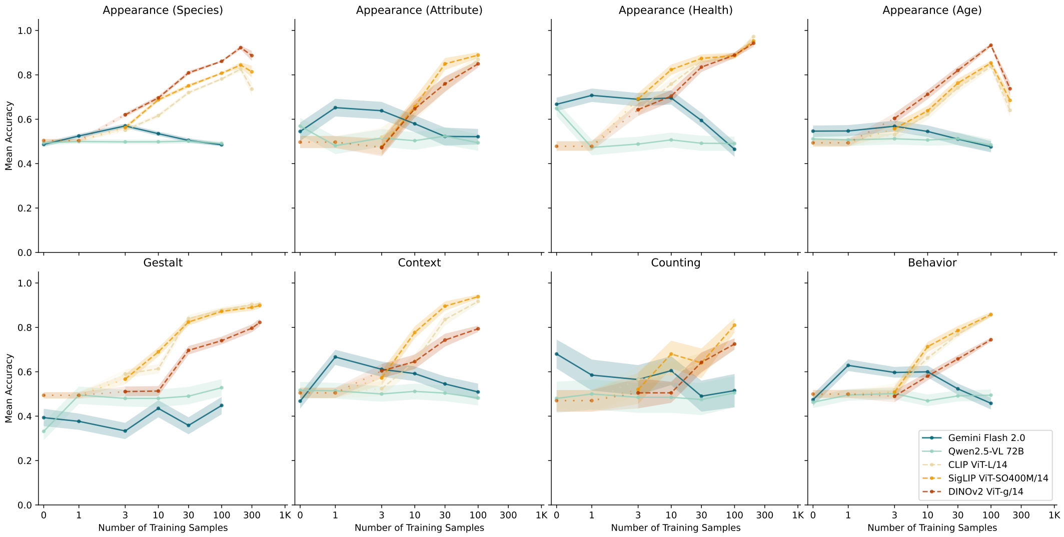

Figure 2: Performance scaling across NeWT's [42] eight task clusters as a function of number of labeled examples. Each panel corresponds to one task *cluster* (species, attributes, health, ages, gestalt, context, counting, behavior; clusters contain more than one task). Lines depict representative multimodal large language models (MLLMs: Gemini Flash 2.0, Qwen2.5-VL 72B) and vision encoders (CLIP ViT-L/14, DINOv2 ViT-g/14, SigLIP ViT-SO400M/14). Shaded regions represent 95% bootstrapped confidence intervals. MLLMs exhibit early performance plateaus compared to sustained improvements seen in vision encoders combined with SVMs as the number of labeled examples. We cannot fit SVMs without at least one labeled example per class; we simulate random chance for 0 and 1 labeled examples.

oversight is significant: models optimized solely for zero- or few-shot benchmarks risk poor alignment with realistic deployments where moderate labeled datasets are both common and crucial.

Our work addresses this critical gap, systematically comparing foundation models (MLLMs) with vision-only methods, explicitly highlighting performance characteristics in the neglected small-data regime.

3 Methodology

To highlight the overlooked small-data regime in AI research, we introduce a rigorous experimental framework comparing MLLMs and vision-only models combined with traditional machine learning approaches specifically within this scenario. Unlike widely-studied zero-shot and few-shot benchmarks, our experiments explicitly target moderate data scales, ranging from tens to thousands of samples, to uncover non-obvious scaling behaviors and limitations of state-of-the-art methods. By providing detailed methodological guidance, we encourage researchers to adopt similar small-data evaluations, facilitating meaningful insights and practical recommendations for future method development.

We evaluate MLLMs and vision-only models with support vector machines (SVMs) as representative paradigms, as both have demonstrated strong performance across diverse visual tasks yet remain insufficiently characterized within the small-data regime. MLLMs have primarily been evaluated on zero-shot or few-shot benchmarks, leaving their performance unclear when moderate quantities of labeled data are available. Conversely, traditional vision encoder-based methods, which explicitly leverage fine-tuning or transfer learning, might exhibit fundamentally different scaling behaviors. Our methodology thus aims to elucidate previously unobserved contrasts and limitations by directly comparing these approaches within this under-explored setting.

Specifically, we evaluate multiple state-of-the-art MLLMs (e.g., Gemini Flash, Qwen2.5-VL) alongside vision encoders (e.g., DINOv2, CLIP variants) paired with SVM-based classifiers on diverse ecological datasets from the NeWT benchmark. We systematically vary the number of labeled examples and apply standard prompting and parsing procedures to rigorously characterize model behaviors and scaling trends. Section 3 contain our prompts and pseudo-code for parsing responses.

Tasks We evaluate models on the NeWT benchmark [42]. NeWT contains 164 ecologically-motivated binary classification tasks, each with 200 to 400 labeled examples. Tasks are grouped into eight clusters: species, attributes, health, ages, gestalt, contexts, counting and behavior. See both Appendix B and the original text for additional details.

Multimodal Large Language Models (MLLMs): We evaluate Gemini Flash 2.0, Gemini Flash 1.5 8B, Qwen2-VL 7B and Qwen2.5-VL 72B.

Vision Encoders with SVMs: We extract features from DINOv2, CLIP, and SigLIP. We test ViT-B, ViT-L, and ViT-H variants. Per NeWT's original methodology, we exclusively use SVMs as our binary classifier on top of dense vision model features; SVM hyperparameters are tuned using `scikit-learn`'s cross-validation grid search [31].

Labeled Examples To analyze performance scaling, we train models on different amounts of labeled examples. We define subsets with sample sizes of 0, 1, 3, 10, 30, 100, 300, and all examples. Labeled examples are sampled uniformly and we ensure that there is at least one example per class when there are two or more examples.

Evaluation For all tasks, we compute bootstrapped confidence intervals by resampling test sets 1,000 times with replacement and reporting the 95% confidence interval. MLLM responses are parsed using deterministic regex-based extraction. If multiple species are listed, we take the first species mentioned.

4 Results

Our experiments reveal distinct performance characteristics between MLLMs and vision-only methods across the small-data spectrum. Specific to ecological computer vision tasks, several notable patterns emerge that challenge conventional assumptions.

4.1 Scaling Data

As shown in Fig. 2, MLLMs and vision-only methods exhibit fundamentally different scaling behaviors as the number of labeled examples increases. MLLMs demonstrate rapid initial gains with very few examples (1-3) but consistently reach performance plateaus after 10-30 examples across most task clusters. In contrast, leveraging SVMs with vision transformers [ViTs; 13] show continuous, near-logarithmic improvement throughout the entire small-data regime, with no evidence of plateauing. This scaling disparity results in a widening performance gap as dataset size increases.

4.2 Scaling Models

Adding parameters to large language models demonstrates consistent improvement [20, 22, 45]. Our analysis reveals a different pattern for vision models in ecological tasks. As Fig. 3a illustrates, increasing computational resources yields diminishing returns compared to simply adding more labeled examples. Even as we scale SigLIP models across model and image sizes from 45 to 700+ GFLOPs, accuracy improvements remain modest, with a 10× increase in labeled samples consistently outperforming a 10× increase in computational capacity. This challenges the dominant "bigger is better" paradigm in recent AI research [26, 44].

Several factors could explain this difference from language model scaling properties. The emergence threshold for vision models might occur at parameter counts beyond our experimental range [12, 15, 36, 37]. Pretraining methodology differences are significant—vision models employ diverse objectives (contrastive, self-supervised, supervised) compared to the converged next-token prediction approach in language. Our findings indicate that for ecological computer vision tasks within the small-data regime, prioritizing data collection provides more reliable performance improvements than scaling computational resources alone.

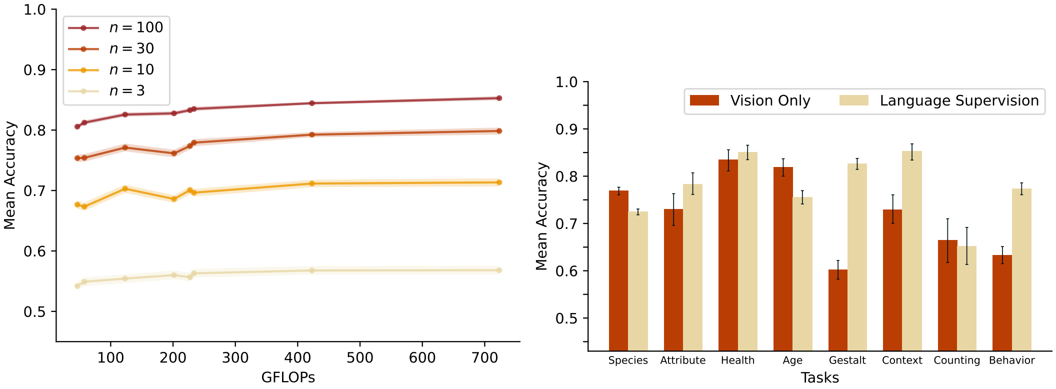

Figure 3: **Left:** Vision model performance with respect to inference FLOPs and number of labeled examples (n). SigLIP [47] released eight different pre-trained transformers with varying model sizes (ViT-B/16, ViT-L/16 and ViT-SO400M/14) and image sizes (224 × 224, 256 × 256, 384 × 384, and 512 × 512); we unify these axes with FLOPs/image. We find that increasing the number of labeled examples is more effective than increasing the model size; a 10× increase in labeled examples outperforms a 10× increase in FLOPs. **Right:** Comparing vision model pre-training on performance across the eight task clusters in NeWT for 30 labeled examples with ViT-L models. Black error bars indicate bootstrapped 95% confidence intervals. Vision-only pre-training [DINOv2; 30] outperforms language-supervised pre-training [CLIP and SigLIP; 32, 47] on 'Species' and 'Age' tasks, both of which are fine-grained classification tasks. We observe that language supervision leads to large improvements on 'Gestalt' and 'Behavior' tasks, both of which require semantic reasoning. These conclusions hold for other numbers of labeled examples; see Appendix C for additional results.

4.3 Pre-Training Supervision

We use these results to reveal distinct performance patterns between vision-only and language-supervised pre-training approaches across ecological task clusters (Fig. 3b). These differences underscore how pre-training objectives fundamentally influence a model's capabilities.

Vision-only pre-training (DINOv2) significantly outperforms language-supervised approaches (CLIP, SigLIP) on fine-grained visual discrimination tasks, specifically 'Species' and 'Age' classification. This advantage likely stems from DINOv2's self-supervised training objective, which builds rich hierarchical representations through local-to-global correspondence without language constraints. Such representations excel at capturing subtle morphological differences crucial for taxonomic identification and age determination tasks.

Conversely, language-supervised pre-training demonstrates substantial advantages in tasks requiring semantic understanding and contextual reasoning, notably in 'Gestalt', 'Context', and 'Behavior' clusters, where models must recognize abstract visual concepts like image quality or animal activities. This suggests that image-text learning provides semantic grounding that pure vision models lack.

These observed differences between pre-training methods are consistent across training set sizes (Appendix C), suggesting fundamental differences in learned representations, which is theoretically supported by recent work in interpreting vision models [35, 39].

5 Conclusion & Future Work

Our systematic evaluation of the small-data regime reveals distinct performance patterns: vision-only systems exhibit sustained improvement while MLLMs demonstrate early performance plateaus beyond 10-30 labeled examples, suggesting that prompting struggles to learn nuanced patterns beyond a critical threshold of examples [21, 48]. Our findings underscore the critical importance of evaluating AI methods explicitly within the small-data regime, an evaluation practice largely overlooked in current work despite its relevance to real applications. By highlighting this evaluation gap, we hope to encourage more comprehensive benchmarking practices that better reflect the diverse data contexts encountered in practice.

References

[1] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. 2

[2] Anthropic. Claude 3.5 sonnet model card addendum, 2024. 2

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[4] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024. 2

[5] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. 1

[6] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021. 2

[7] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1

[8] Lorenzo Brigato and Luca Iocchi. A close look at deep learning with small data. In *2020 25th international conference on pattern recognition (ICPR)*, pages 2490–2497. IEEE, 2021. 2

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[10] Google Deepmind. Gemini 1.5, 2024. 2

[11] Google Deepmind. Gemini 2.0, 2025. 2

[12] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International conference on machine learning*, pages 7480–7512. PMLR, 2023. 4

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[14] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639): 115–118, 2017. 1

[15] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024. 4

[16] Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor Guilherme Turrisi da Costa, Louis Béthune, Zhe Gan, et al. Multimodal autoregressive pre-training of large vision encoders. *arXiv preprint arXiv:2411.14402*, 2024. 2

[17] ZeMing Gong, Austin T Wang, Xiaoliang Huo, Joakim Bruslund Haurum, Scott C Lowe, Graham W Taylor, and Angel X Chang. Clibd: Bridging vision and genomics for biodiversity monitoring at scale. *arXiv preprint arXiv:2405.17537*, 2024. 2

[18] Google. Gemma 3, 2025. 2

[19] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. 1, 2

[20] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 4

[21] Yixing Jiang, Jeremy Andrew Irvin, Ji Hun Wang, Muhammad Ahmed Chaudhry, Jonathan H Chen, and Andrew Y Ng. Many-shot in-context learning in multimodal foundation models. In *ICML 2024 Workshop on In-Context Learning*, 2024. 5

[22] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 4

[23] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021. 2

[24] Ivan Kraljevski, Yong Chul Ju, Dmitrij Ivanov, Constanze Tschöpe, and Matthias Wolff. How to do machine learning with small data?–a review from an industrial perspective. *arXiv preprint arXiv:2311.07126*, 2023. 2

[25] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024. 2

[26] Ian R McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, et al. Inverse scaling: When bigger isn't better. *arXiv preprint arXiv:2306.09479*, 2023. 4

[27] Meta. Llama 3.2, 2024. 2

[28] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024. 2

[29] OpenAI. Gpt-4o system card, 2024. 2

[30] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 2, 5

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 4

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 2, 5

[33] Srikumar Sastry, Subash Khanal, Aayush Dhakal, Adeel Ahmad, and Nathan Jacobs. Taxabind: A unified embedding space for ecological applications. *arXiv preprint arXiv:2411.00683*, 2024. 2

[34] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. BioCLIP: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19412–19424, 2024. 2

[35] Samuel Stevens, Wei-Lun Chao, Tanya Berger-Wolf, and Yu Su. Sparse autoencoders for scientifically rigorous interpretation of vision models, 2025. 5

[36] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 4

[37] Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*, 2024. 4

[38] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022. 2

[39] Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment. *arXiv preprint arXiv:2502.03714*, 2025. 5

[40] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 2

[41] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 1, 2

[42] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Computer Vision and Pattern Recognition*, 2021. 1, 2, 3, 4, 9

[43] Jinjiang Wang, Yulin Ma, Laibin Zhang, Robert X Gao, and Dazhong Wu. Deep learning for smart manufacturing: Methods and applications. *Journal of manufacturing systems*, 48:144–156, 2018. 1

[44] Jason Wei, Najoung Kim, Yi Tay, and Quoc V Le. Inverse scaling can become u-shaped. *arXiv preprint arXiv:2211.02011*, 2022. 4

[45] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022. 4

[46] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024. 1, 2

[47] 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*, pages 11975–11986, 2023. 2, 5

[48] Xingxuan Zhang, Jiansheng Li, Wenjing Chu, Junjia Hai, Renzhe Xu, Yuqing Yang, Shikai Guan, Jiazheng Xu, and Peng Cui. On the out-of-distribution generalization of multimodal large language models. *arXiv preprint arXiv:2402.06599*, 2024. 5

Appendices

1. Appendix A: Methodology details for Section 3.
2. Appendix B: NeWT details for Section 3.
3. Appendix C: Additional results for Section 4.3.
4. Appendix D: All raw results.

A Methodology Details

We evaluate two families of models: (1) multimodal large language models (MLLMs) and (2) vision encoders with machine learning classifiers.

MLLMs: We test the following models via API access: Gemini Flash 2.0, Gemini Flash 1.5 8B, Qwen2-VL 7B and Qwen2.5-VL 72B.

Vision Encoders: We extract image embeddings from DINOv2, CLIP, and SigLIP and include ViT-B, ViT-L, and ViT-H variants.

Image Preprocessing: All images are resized so that the smaller side is 224 pixels, then center cropped to 224×224. We normalize images using the ImageNet mean and standard deviation for all models. No additional augmentations (e.g., cropping, flipping) are applied for inference. Images are not modified before being sent to MLLMs.

Sampling: Subsets of labeled examples are sampled uniformly from the full training split with the following sample sizes: 0, 1, 3, 10, 30, 100, and 300 where applicable.

Prompting and Parsing: All the tasks in NeWT are binary classification tasks. We use the following template: "`What is this a picture of, '{a}' or '{b}'? Respond with your answer in bold.`" where {a} and {b} are replaced with the two classes, with a random order. MLLM responses are parsed using regex-based extraction, using character-based distance to pick the classname closest to whatever bold text is first found in the response.

Classifier Hyperparameters We perform a search over the following hyperparameter distribution:

- C: log-uniform distribution from 10^{-3} to 10^{1}.
- Kernel: one of RBF, linear, sigmoid or cubic
- Kernel coefficient: log-uniform distribution from 10^{-4} to 10^{-3}. Ignored for linear kernel.

We sample 100 models and evaluate with 5-fold cross-validation over the training set.

Bootstrapped Confidence Intervals For all evaluation metrics, we report bootstrapped confidence intervals:

- We resample the test set 1,000 times with replacement.
- The mean accuracy is computed for each resampled test set.
- The 95% confidence interval is reported.

Compute Infrastructure: ViT-based inference is batched on NVIDIA A6000 GPUs to maximize GPU memory efficiency. API-based MLLM inference is conducted on cloud platforms.

B Background on NeWT

Natural World Tasks [NeWT; 42] is a collection of 164 binary classification tasks that go beyond species classification. The tasks are manually curated with a uniform distribution of positive and negative examples (so accuracy is an appropriate metric).

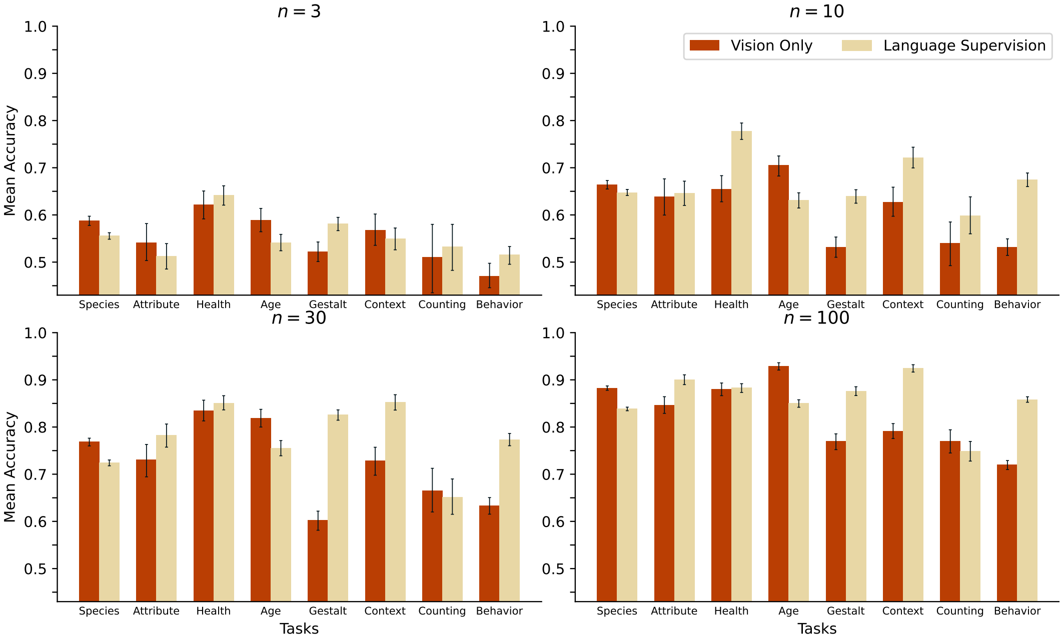

Figure 4: Comparing vision model pre-training on performance across the eight task clusters in NeWT for 3, 10, 30, and 100 training samples with ViT-L models.

C Pre-Training Supervision

Fig. 4 contains results for 3, 10, 30, and 100 training samples, summarized by the ViT's pre-training objective. Vision-only pre-training outperforms vision-language pre-training on 'Species' and 'Age' tasks, while vision-language pre-training outperforms vision-only pre-training on more semantic tasks ('Gestalt', 'Context', and 'Behavior').

D All Results

We include all results in Tables 1 to 18. These results will also be made available in a machine-readable format.

Task Cluster	Task Subcluster	Model	Train	Mean	Confidence Interval
Appearance	Species	Gemini Flash 2.0	0	0.49	$[0.48, 0.50]$
Appearance	Species	Gemini Flash 1.5 8B	0	0.50	$[0.49, 0.51]$
Appearance	Species	Qwen2.5-VL 72B	0	0.49	$[0.48, 0.50]$
Appearance	Species	Qwen2-VL 7B	0	0.50	$[0.49, 0.51]$
Appearance	Attribute	Gemini Flash 2.0	0	0.54	$[0.51, 0.59]$
Appearance	Attribute	Gemini Flash 1.5 8B	0	0.60	$[0.56, 0.64]$
Appearance	Attribute	Qwen2.5-VL 72B	0	0.57	$[0.53, 0.61]$
Appearance	Attribute	Qwen2-VL 7B	0	0.48	$[0.44, 0.52]$
Appearance	Health	Gemini Flash 2.0	0	0.67	$[0.64, 0.70]$
Appearance	Health	Gemini Flash 1.5 8B	0	0.69	$[0.66, 0.72]$
Appearance	Health	Qwen2.5-VL 72B	0	0.65	$[0.61, 0.68]$
Appearance	Health	Qwen2-VL 7B	0	0.61	$[0.58, 0.64]$
Appearance	Age	Gemini Flash 2.0	0	0.55	$[0.52, 0.57]$
Appearance	Age	Gemini Flash 1.5 8B	0	0.50	$[0.47, 0.52]$
Appearance	Age	Qwen2.5-VL 72B	0	0.51	$[0.49, 0.54]$
Appearance	Age	Qwen2-VL 7B	0	0.54	$[0.52, 0.57]$
Gestalt	-	Gemini Flash 2.0	0	0.39	$[0.36, 0.43]$
Gestalt	-	Gemini Flash 1.5 8B	0	0.45	$[0.41, 0.49]$
Gestalt	-	Qwen2.5-VL 72B	0	0.33	$[0.29, 0.37]$
Gestalt	-	Qwen2-VL 7B	0	0.40	$[0.36, 0.44]$
Context	-	Gemini Flash 2.0	0	0.47	$[0.43, 0.50]$
Context	-	Gemini Flash 1.5 8B	0	0.49	$[0.46, 0.53]$
Context	-	Qwen2.5-VL 72B	0	0.52	$[0.48, 0.55]$
Context	-	Qwen2-VL 7B	0	0.53	$[0.50, 0.57]$
Counting	-	Gemini Flash 2.0	0	0.68	$[0.61, 0.74]$
Counting	-	Gemini Flash 1.5 8B	0	0.59	$[0.53, 0.67]$
Counting	-	Qwen2.5-VL 72B	0	0.48	$[0.41, 0.54]$
Counting	-	Qwen2-VL 7B	0	0.64	$[0.56, 0.70]$
Behavior	-	Gemini Flash 2.0	0	0.47	$[0.45, 0.50]$
Behavior	-	Gemini Flash 1.5 8B	0	0.49	$[0.46, 0.51]$
Behavior	-	Qwen2.5-VL 72B	0	0.46	$[0.44, 0.49]$
Behavior	-	Qwen2-VL 7B	0	0.46	$[0.43, 0.48]$

Table 1: All results for 0 training samples.

Task Cluster	Task Subcluster	Model	Train	Mean	Confidence Interval
Appearance	Species	Gemini Flash 2.0	1	0.52	$[0.52, 0.53]$
Appearance	Species	Gemini Flash 1.5 8B	1	0.54	$[0.53, 0.55]$
Appearance	Species	Qwen2.5-VL 72B	1	0.50	$[0.49, 0.51]$
Appearance	Species	Qwen2-VL 7B	1	0.50	$[0.49, 0.51]$
Appearance	Attribute	Gemini Flash 2.0	1	0.65	$[0.61, 0.69]$
Appearance	Attribute	Gemini Flash 1.5 8B	1	0.68	$[0.64, 0.71]$
Appearance	Attribute	Qwen2.5-VL 72B	1	0.48	$[0.44, 0.52]$
Appearance	Attribute	Qwen2-VL 7B	1	0.49	$[0.45, 0.53]$
Appearance	Health	Gemini Flash 2.0	1	0.71	$[0.68, 0.74]$
Appearance	Health	Gemini Flash 1.5 8B	1	0.70	$[0.67, 0.73]$
Appearance	Health	Qwen2.5-VL 72B	1	0.47	$[0.44, 0.51]$
Appearance	Health	Qwen2-VL 7B	1	0.53	$[0.49, 0.56]$
Appearance	Age	Gemini Flash 2.0	1	0.55	$[0.52, 0.57]$
Appearance	Age	Gemini Flash 1.5 8B	1	0.52	$[0.50, 0.55]$
Appearance	Age	Qwen2.5-VL 72B	1	0.51	$[0.48, 0.53]$
Appearance	Age	Qwen2-VL 7B	1	0.50	$[0.47, 0.53]$
Gestalt	-	Gemini Flash 2.0	1	0.38	$[0.34, 0.41]$
Gestalt	-	Gemini Flash 1.5 8B	1	0.42	$[0.38, 0.46]$
Gestalt	-	Qwen2.5-VL 72B	1	0.49	$[0.46, 0.53]$
Gestalt	-	Qwen2-VL 7B	1	0.51	$[0.47, 0.55]$
Context	-	Gemini Flash 2.0	1	0.67	$[0.63, 0.70]$
Context	-	Gemini Flash 1.5 8B	1	0.60	$[0.57, 0.63]$
Context	-	Qwen2.5-VL 72B	1	0.51	$[0.48, 0.55]$
Context	-	Qwen2-VL 7B	1	0.52	$[0.49, 0.56]$
Counting	-	Gemini Flash 2.0	1	0.58	$[0.52, 0.66]$
Counting	-	Gemini Flash 1.5 8B	1	0.62	$[0.56, 0.69]$
Counting	-	Qwen2.5-VL 72B	1	0.50	$[0.43, 0.57]$
Counting	-	Qwen2-VL 7B	1	0.51	$[0.44, 0.57]$
Behavior	-	Gemini Flash 2.0	1	0.63	$[0.60, 0.65]$
Behavior	-	Gemini Flash 1.5 8B	1	0.49	$[0.46, 0.51]$
Behavior	-	Qwen2.5-VL 72B	1	0.50	$[0.47, 0.52]$
Behavior	-	Qwen2-VL 7B	1	0.46	$[0.43, 0.48]$

Table 2: All results for 1 training sample.

Task Cluster	Task Subcluster	Model	Train	Mean	Confidence Interval
Appearance	Species	Gemini Flash 2.0	3	0.57	[0.56, 0.58]
Appearance	Species	Gemini Flash 1.5 8B	3	0.53	[0.52, 0.54]
Appearance	Species	SigLIP ViT-B/16 (512px)	3	0.56	[0.55, 0.57]
Appearance	Species	SigLIP ViT-L/16 (256px)	3	0.56	[0.55, 0.57]
Appearance	Species	SigLIP ViT-L/16 (384px)	3	0.57	[0.56, 0.58]
Appearance	Species	SigLIP ViT-SO400M/14	3	0.56	[0.55, 0.57]
Appearance	Species	SigLIP ViT-SO400M/14 (384px)	3	0.56	[0.56, 0.57]
Appearance	Species	DINOv2 ViT-B/14	3	0.58	[0.57, 0.59]
Appearance	Species	DINOv2 ViT-L/14	3	0.59	[0.58, 0.60]
Appearance	Species	DINOv2 ViT-S/14	3	0.59	[0.58, 0.59]
Appearance	Species	DINOv2 ViT-g/14	3	0.62	[0.61, 0.63]
Appearance	Species	BioCLIP ViT-B/16	3	0.58	[0.57, 0.59]
Appearance	Species	Qwen2.5-VL 72B	3	0.50	[0.49, 0.51]
Appearance	Species	Qwen2-VL 7B	3	0.50	[0.49, 0.51]
Appearance	Species	CLIP ViT-B/16	3	0.54	[0.53, 0.55]
Appearance	Species	CLIP ViT-L/14	3	0.55	[0.54, 0.56]
Appearance	Species	CLIP ViT-L/14 (336px)	3	0.55	[0.54, 0.56]
Appearance	Species	SigLIP ViT-B/16	3	0.54	[0.53, 0.55]
Appearance	Species	SigLIP ViT-B/16 (256px)	3	0.55	[0.54, 0.56]
Appearance	Species	SigLIP ViT-B/16 (384px)	3	0.55	[0.54, 0.56]
Appearance	Attribute	Gemini Flash 2.0	3	0.64	[0.60, 0.68]
Appearance	Attribute	Gemini Flash 1.5 8B	3	0.58	[0.54, 0.62]
Appearance	Attribute	SigLIP ViT-B/16 (512px)	3	0.46	[0.42, 0.50]
Appearance	Attribute	SigLIP ViT-L/16 (256px)	3	0.50	[0.46, 0.54]
Appearance	Attribute	SigLIP ViT-L/16 (384px)	3	0.51	[0.48, 0.55]
Appearance	Attribute	SigLIP ViT-SO400M/14	3	0.48	[0.43, 0.52]
Appearance	Attribute	SigLIP ViT-SO400M/14 (384px)	3	0.49	[0.45, 0.53]
Appearance	Attribute	DINOv2 ViT-B/14	3	0.50	[0.47, 0.54]
Appearance	Attribute	DINOv2 ViT-L/14	3	0.54	[0.50, 0.58]
Appearance	Attribute	DINOv2 ViT-S/14	3	0.50	[0.46, 0.54]
Appearance	Attribute	DINOv2 ViT-g/14	3	0.47	[0.43, 0.51]
Appearance	Attribute	BioCLIP ViT-B/16	3	0.52	[0.49, 0.57]
Appearance	Attribute	Qwen2.5-VL 72B	3	0.52	[0.47, 0.56]
Appearance	Attribute	Qwen2-VL 7B	3	0.53	[0.48, 0.57]
Appearance	Attribute	CLIP ViT-B/16	3	0.50	[0.46, 0.54]
Appearance	Attribute	CLIP ViT-L/14	3	0.52	[0.48, 0.56]
Appearance	Attribute	CLIP ViT-L/14 (336px)	3	0.48	[0.44, 0.52]
Appearance	Attribute	SigLIP ViT-B/16	3	0.47	[0.43, 0.50]
Appearance	Attribute	SigLIP ViT-B/16 (256px)	3	0.47	[0.43, 0.51]
Appearance	Attribute	SigLIP ViT-B/16 (384px)	3	0.48	[0.44, 0.53]

Table 3: All results for 3 training samples for ‘Species’ and ‘Attribute’ tasks.

Task Cluster	Task Subcluster	Model	Train	Mean	Confidence Interval
Appearance	Health	Gemini Flash 2.0	3	0.69	$[0.66, 0.72]$
Appearance	Health	Gemini Flash 1.5 8B	3	0.71	$[0.68, 0.74]$
Appearance	Health	SigLIP ViT-B/16 (512px)	3	0.59	$[0.56, 0.62]$
Appearance	Health	SigLIP ViT-L/16 (256px)	3	0.64	$[0.61, 0.67]$
Appearance	Health	SigLIP ViT-L/16 (384px)	3	0.69	$[0.66, 0.72]$
Appearance	Health	SigLIP ViT-SO400M/14	3	0.69	$[0.67, 0.72]$
Appearance	Health	SigLIP ViT-SO400M/14 (384px)	3	0.73	$[0.70, 0.76]$
Appearance	Health	DINOv2 ViT-B/14	3	0.64	$[0.62, 0.67]$
Appearance	Health	DINOv2 ViT-L/14	3	0.62	$[0.59, 0.65]$
Appearance	Health	DINOv2 ViT-S/14	3	0.57	$[0.54, 0.60]$
Appearance	Health	DINOv2 ViT-g/14	3	0.64	$[0.61, 0.67]$
Appearance	Health	BioCLIP ViT-B/16	3	0.62	$[0.59, 0.65]$
Appearance	Health	Qwen2.5-VL 72B	3	0.49	$[0.46, 0.52]$
Appearance	Health	Qwen2-VL 7B	3	0.47	$[0.44, 0.51]$
Appearance	Health	CLIP ViT-B/16	3	0.61	$[0.58, 0.64]$
Appearance	Health	CLIP ViT-L/14	3	0.64	$[0.61, 0.67]$
Appearance	Health	CLIP ViT-L/14 (336px)	3	0.57	$[0.55, 0.60]$
Appearance	Health	SigLIP ViT-B/16	3	0.58	$[0.55, 0.61]$
Appearance	Health	SigLIP ViT-B/16 (256px)	3	0.58	$[0.55, 0.61]$
Appearance	Health	SigLIP ViT-B/16 (384px)	3	0.62	$[0.59, 0.65]$
Appearance	Age	Gemini Flash 2.0	3	0.57	$[0.54, 0.59]$
Appearance	Age	Gemini Flash 1.5 8B	3	0.50	$[0.47, 0.52]$
Appearance	Age	SigLIP ViT-B/16 (512px)	3	0.54	$[0.52, 0.57]$
Appearance	Age	SigLIP ViT-L/16 (256px)	3	0.55	$[0.53, 0.58]$
Appearance	Age	SigLIP ViT-L/16 (384px)	3	0.57	$[0.54, 0.59]$
Appearance	Age	SigLIP ViT-SO400M/14	3	0.56	$[0.53, 0.58]$
Appearance	Age	SigLIP ViT-SO400M/14 (384px)	3	0.57	$[0.55, 0.59]$
Appearance	Age	DINOv2 ViT-B/14	3	0.58	$[0.55, 0.60]$
Appearance	Age	DINOv2 ViT-L/14	3	0.59	$[0.56, 0.61]$
Appearance	Age	DINOv2 ViT-S/14	3	0.63	$[0.60, 0.65]$
Appearance	Age	DINOv2 ViT-g/14	3	0.60	$[0.58, 0.63]$
Appearance	Age	BioCLIP ViT-B/16	3	0.53	$[0.50, 0.55]$
Appearance	Age	Qwen2.5-VL 72B	3	0.51	$[0.48, 0.54]$
Appearance	Age	Qwen2-VL 7B	3	0.50	$[0.47, 0.52]$
Appearance	Age	CLIP ViT-B/16	3	0.53	$[0.51, 0.56]$
Appearance	Age	CLIP ViT-L/14	3	0.53	$[0.51, 0.55]$
Appearance	Age	CLIP ViT-L/14 (336px)	3	0.53	$[0.51, 0.55]$
Appearance	Age	SigLIP ViT-B/16	3	0.53	$[0.51, 0.56]$
Appearance	Age	SigLIP ViT-B/16 (256px)	3	0.56	$[0.53, 0.58]$
Appearance	Age	SigLIP ViT-B/16 (384px)	3	0.55	$[0.53, 0.57]$

Table 4: All results for 3 training samples for ‘Health’ and ‘Age’ tasks.

Task Cluster	Task Subcluster	Model	Train	Mean	Confidence Interval
Gestalt	-	Gemini Flash 2.0	3	0.33	$[0.29, 0.37]$
Gestalt	-	Gemini Flash 1.5 8B	3	0.45	$[0.41, 0.49]$
Gestalt	-	SigLIP ViT-B/16 (512px)	3	0.61	$[0.59, 0.63]$
Gestalt	-	SigLIP ViT-L/16 (256px)	3	0.57	$[0.55, 0.59]$
Gestalt	-	SigLIP ViT-L/16 (384px)	3	0.56	$[0.54, 0.58]$
Gestalt	-	SigLIP ViT-SO400M/14	3	0.57	$[0.55, 0.59]$
Gestalt	-	SigLIP ViT-SO400M/14 (384px)	3	0.57	$[0.54, 0.59]$
Gestalt	-	DINOv2 ViT-B/14	3	0.51	$[0.49, 0.53]$
Gestalt	-	DINOv2 ViT-L/14	3	0.52	$[0.50, 0.54]$
Gestalt	-	DINOv2 ViT-S/14	3	0.51	$[0.49, 0.53]$
Gestalt	-	DINOv2 ViT-g/14	3	0.51	$[0.49, 0.53]$
Gestalt	-	BioCLIP ViT-B/16	3	0.51	$[0.49, 0.53]$
Gestalt	-	Qwen2.5-VL 72B	3	0.48	$[0.44, 0.52]$
Gestalt	-	Qwen2-VL 7B	3	0.51	$[0.47, 0.55]$
Gestalt	-	CLIP ViT-B/16	3	0.57	$[0.55, 0.59]$
Gestalt	-	CLIP ViT-L/14	3	0.59	$[0.57, 0.61]$
Gestalt	-	CLIP ViT-L/14 (336px)	3	0.53	$[0.51, 0.56]$
Gestalt	-	SigLIP ViT-B/16	3	0.58	$[0.56, 0.60]$
Gestalt	-	SigLIP ViT-B/16 (256px)	3	0.59	$[0.57, 0.61]$
Gestalt	-	SigLIP ViT-B/16 (384px)	3	0.59	$[0.57, 0.61]$
Context	-	Gemini Flash 2.0	3	0.61	$[0.58, 0.65]$
Context	-	Gemini Flash 1.5 8B	3	0.53	$[0.50, 0.57]$
Context	-	SigLIP ViT-B/16 (512px)	3	0.57	$[0.54, 0.60]$
Context	-	SigLIP ViT-L/16 (256px)	3	0.58	$[0.54, 0.61]$
Context	-	SigLIP ViT-L/16 (384px)	3	0.56	$[0.52, 0.59]$
Context	-	SigLIP ViT-SO400M/14	3	0.57	$[0.54, 0.60]$
Context	-	SigLIP ViT-SO400M/14 (384px)	3	0.55	$[0.52, 0.59]$
Context	-	DINOv2 ViT-B/14	3	0.61	$[0.58, 0.64]$
Context	-	DINOv2 ViT-L/14	3	0.57	$[0.53, 0.60]$
Context	-	DINOv2 ViT-S/14	3	0.54	$[0.50, 0.57]$
Context	-	DINOv2 ViT-g/14	3	0.60	$[0.57, 0.64]$
Context	-	BioCLIP ViT-B/16	3	0.52	$[0.48, 0.55]$
Context	-	Qwen2.5-VL 72B	3	0.50	$[0.46, 0.53]$
Context	-	Qwen2-VL 7B	3	0.48	$[0.45, 0.52]$
Context	-	CLIP ViT-B/16	3	0.55	$[0.51, 0.58]$
Context	-	CLIP ViT-L/14	3	0.52	$[0.49, 0.56]$
Context	-	CLIP ViT-L/14 (336px)	3	0.54	$[0.50, 0.57]$
Context	-	SigLIP ViT-B/16	3	0.58	$[0.55, 0.62]$
Context	-	SigLIP ViT-B/16 (256px)	3	0.60	$[0.56, 0.63]$
Context	-	SigLIP ViT-B/16 (384px)	3	0.58	$[0.55, 0.61]$

Table 5: All results for 3 training samples for ‘Gestalt’ and ‘Context’ tasks.

Task Cluster	Task Subcluster	Model	Train	Mean	Confidence Interval
Counting	-	Gemini Flash 2.0	3	0.56	[0.50, 0.63]
Counting	-	Gemini Flash 1.5 8B	3	0.61	[0.54, 0.68]
Counting	-	SigLIP ViT-B/16 (512px)	3	0.49	[0.42, 0.56]
Counting	-	SigLIP ViT-L/16 (256px)	3	0.50	[0.43, 0.57]
Counting	-	SigLIP ViT-L/16 (384px)	3	0.50	[0.43, 0.56]
Counting	-	SigLIP ViT-SO400M/14	3	0.52	[0.45, 0.59]
Counting	-	SigLIP ViT-SO400M/14 (384px)	3	0.52	[0.45, 0.59]
Counting	-	DINOv2 ViT-B/14	3	0.52	[0.45, 0.58]
Counting	-	DINOv2 ViT-L/14	3	0.51	[0.44, 0.57]
Counting	-	DINOv2 ViT-S/14	3	0.53	[0.46, 0.60]
Counting	-	DINOv2 ViT-g/14	3	0.51	[0.44, 0.57]
Counting	-	BioCLIP ViT-B/16	3	0.49	[0.42, 0.56]
Counting	-	Qwen2.5-VL 72B	3	0.48	[0.41, 0.56]
Counting	-	Qwen2-VL 7B	3	0.52	[0.44, 0.59]
Counting	-	CLIP ViT-B/16	3	0.51	[0.45, 0.58]
Counting	-	CLIP ViT-L/14	3	0.56	[0.50, 0.64]
Counting	-	CLIP ViT-L/14 (336px)	3	0.48	[0.41, 0.55]
Counting	-	SigLIP ViT-B/16	3	0.50	[0.43, 0.57]
Counting	-	SigLIP ViT-B/16 (256px)	3	0.49	[0.42, 0.56]
Counting	-	SigLIP ViT-B/16 (384px)	3	0.49	[0.43, 0.56]
Behavior	-	Gemini Flash 2.0	3	0.60	[0.57, 0.62]
Behavior	-	Gemini Flash 1.5 8B	3	0.51	[0.49, 0.54]
Behavior	-	SigLIP ViT-B/16 (512px)	3	0.51	[0.48, 0.54]
Behavior	-	SigLIP ViT-L/16 (256px)	3	0.51	[0.48, 0.54]
Behavior	-	SigLIP ViT-L/16 (384px)	3	0.51	[0.49, 0.54]
Behavior	-	SigLIP ViT-SO400M/14	3	0.51	[0.48, 0.53]
Behavior	-	SigLIP ViT-SO400M/14 (384px)	3	0.52	[0.50, 0.55]
Behavior	-	DINOv2 ViT-B/14	3	0.47	[0.44, 0.49]
Behavior	-	DINOv2 ViT-L/14	3	0.47	[0.44, 0.50]
Behavior	-	DINOv2 ViT-S/14	3	0.50	[0.48, 0.53]
Behavior	-	DINOv2 ViT-g/14	3	0.49	[0.46, 0.51]
Behavior	-	BioCLIP ViT-B/16	3	0.50	[0.48, 0.53]
Behavior	-	Qwen2.5-VL 72B	3	0.50	[0.47, 0.53]
Behavior	-	Qwen2-VL 7B	3	0.52	[0.49, 0.54]
Behavior	-	CLIP ViT-B/16	3	0.50	[0.47, 0.52]
Behavior	-	CLIP ViT-L/14	3	0.52	[0.49, 0.54]
Behavior	-	CLIP ViT-L/14 (336px)	3	0.51	[0.49, 0.54]
Behavior	-	SigLIP ViT-B/16	3	0.48	[0.46, 0.51]
Behavior	-	SigLIP ViT-B/16 (256px)	3	0.49	[0.47, 0.52]
Behavior	-	SigLIP ViT-B/16 (384px)	3	0.51	[0.48, 0.53]

Table 6: All results for 3 training samples for ‘Counting’ and ‘Behavior’ tasks.

Task Cluster	Task Subcluster	Model	Train	Mean	Confidence Interval
Appearance	Species	Gemini Flash 2.0	10	0.53	$[0.53, 0.54]$
Appearance	Species	Gemini Flash 1.5 8B	10	0.50	$[0.49, 0.52]$
Appearance	Species	SigLIP ViT-B/16 (512px)	10	0.69	$[0.68, 0.70]$
Appearance	Species	SigLIP ViT-L/16 (256px)	10	0.68	$[0.67, 0.69]$
Appearance	Species	SigLIP ViT-L/16 (384px)	10	0.70	$[0.69, 0.70]$
Appearance	Species	SigLIP ViT-SO400M/14	10	0.69	$[0.68, 0.70]$
Appearance	Species	SigLIP ViT-SO400M/14 (384px)	10	0.70	$[0.69, 0.71]$
Appearance	Species	DINOv2 ViT-B/14	10	0.66	$[0.65, 0.67]$
Appearance	Species	DINOv2 ViT-L/14	10	0.66	$[0.66, 0.67]$
Appearance	Species	DINOv2 ViT-S/14	10	0.68	$[0.67, 0.69]$
Appearance	Species	DINOv2 ViT-g/14	10	0.70	$[0.69, 0.70]$
Appearance	Species	BioCLIP ViT-B/16	10	0.69	$[0.68, 0.70]$
Appearance	Species	Qwen2.5-VL 72B	10	0.50	$[0.49, 0.51]$
Appearance	Species	Qwen2-VL 7B	10	0.50	$[0.49, 0.51]$
Appearance	Species	CLIP ViT-B/16	10	0.59	$[0.58, 0.60]$
Appearance	Species	CLIP ViT-L/14	10	0.62	$[0.61, 0.63]$
Appearance	Species	CLIP ViT-L/14 (336px)	10	0.62	$[0.61, 0.63]$
Appearance	Species	SigLIP ViT-B/16	10	0.67	$[0.66, 0.68]$
Appearance	Species	SigLIP ViT-B/16 (256px)	10	0.67	$[0.66, 0.68]$
Appearance	Species	SigLIP ViT-B/16 (384px)	10	0.69	$[0.68, 0.70]$
Appearance	Attribute	Gemini Flash 2.0	10	0.58	$[0.54, 0.62]$
Appearance	Attribute	Gemini Flash 1.5 8B	10	0.54	$[0.50, 0.58]$
Appearance	Attribute	SigLIP ViT-B/16 (512px)	10	0.63	$[0.59, 0.66]$
Appearance	Attribute	SigLIP ViT-L/16 (256px)	10	0.66	$[0.62, 0.69]$
Appearance	Attribute	SigLIP ViT-L/16 (384px)	10	0.67	$[0.64, 0.70]$
Appearance	Attribute	SigLIP ViT-SO400M/14	10	0.66	$[0.62, 0.69]$
Appearance	Attribute	SigLIP ViT-SO400M/14 (384px)	10	0.69	$[0.66, 0.73]$
Appearance	Attribute	DINOv2 ViT-B/14	10	0.64	$[0.60, 0.68]$
Appearance	Attribute	DINOv2 ViT-L/14	10	0.64	$[0.60, 0.67]$
Appearance	Attribute	DINOv2 ViT-S/14	10	0.65	$[0.61, 0.69]$
Appearance	Attribute	DINOv2 ViT-g/14	10	0.65	$[0.61, 0.68]$
Appearance	Attribute	BioCLIP ViT-B/16	10	0.67	$[0.64, 0.71]$
Appearance	Attribute	Qwen2.5-VL 72B	10	0.50	$[0.46, 0.55]$
Appearance	Attribute	Qwen2-VL 7B	10	0.52	$[0.48, 0.56]$
Appearance	Attribute	CLIP ViT-B/16	10	0.58	$[0.54, 0.62]$
Appearance	Attribute	CLIP ViT-L/14	10	0.64	$[0.60, 0.67]$
Appearance	Attribute	CLIP ViT-L/14 (336px)	10	0.67	$[0.64, 0.71]$
Appearance	Attribute	SigLIP ViT-B/16	10	0.64	$[0.60, 0.67]$
Appearance	Attribute	SigLIP ViT-B/16 (256px)	10	0.63	$[0.59, 0.67]$
Appearance	Attribute	SigLIP ViT-B/16 (384px)	10	0.61	$[0.57, 0.65]$

Table 7: All results for 10 training samples for ‘Species’ and ‘Attribute’ tasks.

Task Cluster	Task Subcluster	Model	Train	Mean	Confidence Interval
Appearance	Health	Gemini Flash 2.0	10	0.70	$[0.66, 0.73]$
Appearance	Health	Gemini Flash 1.5 8B	10	0.55	$[0.52, 0.59]$
Appearance	Health	SigLIP ViT-B/16 (512px)	10	0.81	$[0.79, 0.84]$
Appearance	Health	SigLIP ViT-L/16 (256px)	10	0.80	$[0.77, 0.82]$
Appearance	Health	SigLIP ViT-L/16 (384px)	10	0.83	$[0.81, 0.85]$
Appearance	Health	SigLIP ViT-SO400M/14	10	0.82	$[0.80, 0.85]$
Appearance	Health	SigLIP ViT-SO400M/14 (384px)	10	0.87	$[0.85, 0.89]$
Appearance	Health	DINOv2 ViT-B/14	10	0.68	$[0.65, 0.71]$
Appearance	Health	DINOv2 ViT-L/14	10	0.65	$[0.63, 0.68]$
Appearance	Health	DINOv2 ViT-S/14	10	0.73	$[0.70, 0.75]$
Appearance	Health	DINOv2 ViT-g/14	10	0.70	$[0.67, 0.73]$
Appearance	Health	BioCLIP ViT-B/16	10	0.69	$[0.66, 0.72]$
Appearance	Health	Qwen2.5-VL 72B	10	0.51	$[0.47, 0.54]$
Appearance	Health	Qwen2-VL 7B	10	0.52	$[0.49, 0.56]$
Appearance	Health	CLIP ViT-B/16	10	0.74	$[0.71, 0.77]$
Appearance	Health	CLIP ViT-L/14	10	0.76	$[0.73, 0.79]$
Appearance	Health	CLIP ViT-L/14 (336px)	10	0.65	$[0.62, 0.68]$
Appearance	Health	SigLIP ViT-B/16	10	0.78	$[0.75, 0.80]$
Appearance	Health	SigLIP ViT-B/16 (256px)	10	0.79	$[0.76, 0.81]$
Appearance	Health	SigLIP ViT-B/16 (384px)	10	0.82	$[0.80, 0.85]$
Appearance	Age	Gemini Flash 2.0	10	0.55	$[0.52, 0.57]$
Appearance	Age	Gemini Flash 1.5 8B	10	0.49	$[0.46, 0.52]$
Appearance	Age	SigLIP ViT-B/16 (512px)	10	0.66	$[0.64, 0.68]$
Appearance	Age	SigLIP ViT-L/16 (256px)	10	0.64	$[0.62, 0.66]$
Appearance	Age	SigLIP ViT-L/16 (384px)	10	0.66	$[0.64, 0.68]$
Appearance	Age	SigLIP ViT-SO400M/14	10	0.64	$[0.61, 0.66]$
Appearance	Age	SigLIP ViT-SO400M/14 (384px)	10	0.68	$[0.65, 0.70]$
Appearance	Age	DINOv2 ViT-B/14	10	0.69	$[0.67, 0.71]$
Appearance	Age	DINOv2 ViT-L/14	10	0.70	$[0.68, 0.73]$
Appearance	Age	DINOv2 ViT-S/14	10	0.73	$[0.71, 0.75]$
Appearance	Age	DINOv2 ViT-g/14	10	0.71	$[0.69, 0.73]$
Appearance	Age	BioCLIP ViT-B/16	10	0.62	$[0.60, 0.64]$
Appearance	Age	Qwen2.5-VL 72B	10	0.51	$[0.48, 0.53]$
Appearance	Age	Qwen2-VL 7B	10	0.51	$[0.48, 0.53]$
Appearance	Age	CLIP ViT-B/16	10	0.59	$[0.56, 0.61]$
Appearance	Age	CLIP ViT-L/14	10	0.62	$[0.60, 0.65]$
Appearance	Age	CLIP ViT-L/14 (336px)	10	0.63	$[0.61, 0.65]$
Appearance	Age	SigLIP ViT-B/16	10	0.62	$[0.60, 0.65]$
Appearance	Age	SigLIP ViT-B/16 (256px)	10	0.63	$[0.60, 0.65]$
Appearance	Age	SigLIP ViT-B/16 (384px)	10	0.67	$[0.64, 0.69]$

Table 8: All results for 10 training samples for ‘Health’ and ‘Age’ tasks.

Task Cluster	Task Subcluster	Model	Train	Mean	Confidence Interval
Gestalt	-	Gemini Flash 2.0	10	0.43	$[0.40, 0.48]$
Gestalt	-	Gemini Flash 1.5 8B	10	0.49	$[0.45, 0.53]$
Gestalt	-	SigLIP ViT-B/16 (512px)	10	0.71	$[0.69, 0.73]$
Gestalt	-	SigLIP ViT-L/16 (256px)	10	0.66	$[0.64, 0.69]$
Gestalt	-	SigLIP ViT-L/16 (384px)	10	0.69	$[0.67, 0.71]$
Gestalt	-	SigLIP ViT-SO400M/14	10	0.69	$[0.67, 0.71]$
Gestalt	-	SigLIP ViT-SO400M/14 (384px)	10	0.70	$[0.68, 0.72]$
Gestalt	-	DINOv2 ViT-B/14	10	0.51	$[0.48, 0.53]$
Gestalt	-	DINOv2 ViT-L/14	10	0.53	$[0.51, 0.55]$
Gestalt	-	DINOv2 ViT-S/14	10	0.62	$[0.60, 0.64]$
Gestalt	-	DINOv2 ViT-g/14	10	0.51	$[0.49, 0.53]$
Gestalt	-	BioCLIP ViT-B/16	10	0.54	$[0.52, 0.57]$
Gestalt	-	Qwen2.5-VL 72B	10	0.48	$[0.44, 0.52]$
Gestalt	-	Qwen2-VL 7B	10	0.48	$[0.44, 0.52]$
Gestalt	-	CLIP ViT-B/16	10	0.79	$[0.77, 0.80]$
Gestalt	-	CLIP ViT-L/14	10	0.61	$[0.59, 0.64]$
Gestalt	-	CLIP ViT-L/14 (336px)	10	0.71	$[0.69, 0.73]$
Gestalt	-	SigLIP ViT-B/16	10	0.71	$[0.69, 0.73]$
Gestalt	-	SigLIP ViT-B/16 (256px)	10	0.67	$[0.65, 0.68]$
Gestalt	-	SigLIP ViT-B/16 (384px)	10	0.73	$[0.71, 0.75]$
Context	-	Gemini Flash 2.0	10	0.59	$[0.55, 0.63]$
Context	-	Gemini Flash 1.5 8B	10	0.56	$[0.53, 0.60]$
Context	-	SigLIP ViT-B/16 (512px)	10	0.78	$[0.75, 0.81]$
Context	-	SigLIP ViT-L/16 (256px)	10	0.80	$[0.77, 0.82]$
Context	-	SigLIP ViT-L/16 (384px)	10	0.75	$[0.73, 0.78]$
Context	-	SigLIP ViT-SO400M/14	10	0.78	$[0.75, 0.81]$
Context	-	SigLIP ViT-SO400M/14 (384px)	10	0.77	$[0.74, 0.80]$
Context	-	DINOv2 ViT-B/14	10	0.60	$[0.57, 0.63]$
Context	-	DINOv2 ViT-L/14	10	0.63	$[0.59, 0.66]$
Context	-	DINOv2 ViT-S/14	10	0.56	$[0.52, 0.59]$
Context	-	DINOv2 ViT-g/14	10	0.65	$[0.61, 0.68]$
Context	-	BioCLIP ViT-B/16	10	0.57	$[0.53, 0.60]$
Context	-	Qwen2.5-VL 72B	10	0.51	$[0.48, 0.54]$
Context	-	Qwen2-VL 7B	10	0.49	$[0.46, 0.53]$
Context	-	CLIP ViT-B/16	10	0.67	$[0.64, 0.70]$
Context	-	CLIP ViT-L/14	10	0.65	$[0.61, 0.68]$
Context	-	CLIP ViT-L/14 (336px)	10	0.71	$[0.68, 0.74]$
Context	-	SigLIP ViT-B/16	10	0.75	$[0.72, 0.78]$
Context	-	SigLIP ViT-B/16 (256px)	10	0.78	$[0.75, 0.81]$
Context	-	SigLIP ViT-B/16 (384px)	10	0.80	$[0.77, 0.82]$

Table 9: All results for 10 training samples for ‘Gestalt’ and ‘Context’ tasks.

Task Cluster	Task Subcluster	Model	Train	Mean	Confidence Interval
Counting	-	Gemini Flash 2.0	10	0.60	[0.53, 0.68]
Counting	-	Gemini Flash 1.5 8B	10	0.51	[0.44, 0.57]
Counting	-	SigLIP ViT-B/16 (512px)	10	0.56	[0.49, 0.62]
Counting	-	SigLIP ViT-L/16 (256px)	10	0.61	[0.55, 0.68]
Counting	-	SigLIP ViT-L/16 (384px)	10	0.66	[0.58, 0.71]
Counting	-	SigLIP ViT-SO400M/14	10	0.68	[0.61, 0.74]
Counting	-	SigLIP ViT-SO400M/14 (384px)	10	0.69	[0.61, 0.74]
Counting	-	DINOv2 ViT-B/14	10	0.49	[0.44, 0.54]
Counting	-	DINOv2 ViT-L/14	10	0.54	[0.49, 0.59]
Counting	-	DINOv2 ViT-S/14	10	0.50	[0.45, 0.55]
Counting	-	DINOv2 ViT-g/14	10	0.51	[0.46, 0.56]
Counting	-	BioCLIP ViT-B/16	10	0.56	[0.51, 0.60]
Counting	-	Qwen2.5-VL 72B	10	0.48	[0.41, 0.56]
Counting	-	Qwen2-VL 7B	10	0.52	[0.45, 0.59]
Counting	-	CLIP ViT-B/16	10	0.53	[0.47, 0.57]
Counting	-	CLIP ViT-L/14	10	0.59	[0.54, 0.64]
Counting	-	CLIP ViT-L/14 (336px)	10	0.53	[0.48, 0.58]
Counting	-	SigLIP ViT-B/16	10	0.60	[0.54, 0.67]
Counting	-	SigLIP ViT-B/16 (256px)	10	0.61	[0.55, 0.68]
Counting	-	SigLIP ViT-B/16 (384px)	10	0.64	[0.56, 0.70]
Behavior	-	Gemini Flash 2.0	10	0.60	[0.57, 0.63]
Behavior	-	Gemini Flash 1.5 8B	10	0.46	[0.44, 0.49]
Behavior	-	SigLIP ViT-B/16 (512px)	10	0.71	[0.69, 0.74]
Behavior	-	SigLIP ViT-L/16 (256px)	10	0.70	[0.68, 0.72]
Behavior	-	SigLIP ViT-L/16 (384px)	10	0.79	[0.77, 0.81]
Behavior	-	SigLIP ViT-SO400M/14	10	0.71	[0.69, 0.74]
Behavior	-	SigLIP ViT-SO400M/14 (384px)	10	0.76	[0.73, 0.78]
Behavior	-	DINOv2 ViT-B/14	10	0.56	[0.54, 0.58]
Behavior	-	DINOv2 ViT-L/14	10	0.53	[0.51, 0.55]
Behavior	-	DINOv2 ViT-S/14	10	0.57	[0.56, 0.59]
Behavior	-	DINOv2 ViT-g/14	10	0.58	[0.56, 0.60]
Behavior	-	BioCLIP ViT-B/16	10	0.55	[0.54, 0.57]
Behavior	-	Qwen2.5-VL 72B	10	0.47	[0.44, 0.49]
Behavior	-	Qwen2-VL 7B	10	0.52	[0.49, 0.54]
Behavior	-	CLIP ViT-B/16	10	0.58	[0.56, 0.59]
Behavior	-	CLIP ViT-L/14	10	0.66	[0.64, 0.68]
Behavior	-	CLIP ViT-L/14 (336px)	10	0.66	[0.64, 0.68]
Behavior	-	SigLIP ViT-B/16	10	0.66	[0.63, 0.68]
Behavior	-	SigLIP ViT-B/16 (256px)	10	0.65	[0.62, 0.67]
Behavior	-	SigLIP ViT-B/16 (384px)	10	0.71	[0.69, 0.74]

Table 10: All results for 10 training samples for ‘Counting’ and ‘Behavior’ tasks.

Task Cluster	Task Subcluster	Model	Train	Mean	Confidence Interval
Appearance	Species	Gemini Flash 2.0	30	0.50	[0.49, 0.51]
Appearance	Species	Gemini Flash 1.5 8B	30	0.50	[0.49, 0.51]
Appearance	Species	SigLIP ViT-B/16 (512px)	30	0.74	[0.73, 0.75]
Appearance	Species	SigLIP ViT-L/16 (256px)	30	0.73	[0.72, 0.74]
Appearance	Species	SigLIP ViT-L/16 (384px)	30	0.76	[0.75, 0.76]
Appearance	Species	SigLIP ViT-SO400M/14	30	0.75	[0.74, 0.76]
Appearance	Species	SigLIP ViT-SO400M/14 (384px)	30	0.77	[0.76, 0.78]
Appearance	Species	DINOv2 ViT-B/14	30	0.78	[0.78, 0.79]
Appearance	Species	DINOv2 ViT-L/14	30	0.77	[0.76, 0.78]
Appearance	Species	DINOv2 ViT-S/14	30	0.77	[0.76, 0.78]
Appearance	Species	DINOv2 ViT-g/14	30	0.81	[0.80, 0.82]
Appearance	Species	BioCLIP ViT-B/16	30	0.79	[0.78, 0.80]
Appearance	Species	Qwen2.5-VL 72B	30	0.50	[0.49, 0.51]
Appearance	Species	Qwen2-VL 7B	30	0.50	[0.49, 0.51]
Appearance	Species	CLIP ViT-B/16	30	0.70	[0.69, 0.71]
Appearance	Species	CLIP ViT-L/14	30	0.72	[0.71, 0.73]
Appearance	Species	CLIP ViT-L/14 (336px)	30	0.72	[0.71, 0.73]
Appearance	Species	SigLIP ViT-B/16	30	0.73	[0.72, 0.73]
Appearance	Species	SigLIP ViT-B/16 (256px)	30	0.72	[0.71, 0.73]
Appearance	Species	SigLIP ViT-B/16 (384px)	30	0.74	[0.73, 0.74]
Appearance	Attribute	Gemini Flash 2.0	30	0.52	[0.48, 0.56]
Appearance	Attribute	Gemini Flash 1.5 8B	30	0.51	[0.47, 0.55]
Appearance	Attribute	SigLIP ViT-B/16 (512px)	30	0.83	[0.80, 0.86]
Appearance	Attribute	SigLIP ViT-L/16 (256px)	30	0.81	[0.78, 0.84]
Appearance	Attribute	SigLIP ViT-L/16 (384px)	30	0.83	[0.80, 0.85]
Appearance	Attribute	SigLIP ViT-SO400M/14	30	0.85	[0.82, 0.88]
Appearance	Attribute	SigLIP ViT-SO400M/14 (384px)	30	0.82	[0.79, 0.85]
Appearance	Attribute	DINOv2 ViT-B/14	30	0.80	[0.77, 0.83]
Appearance	Attribute	DINOv2 ViT-L/14	30	0.73	[0.69, 0.76]
Appearance	Attribute	DINOv2 ViT-S/14	30	0.80	[0.76, 0.83]
Appearance	Attribute	DINOv2 ViT-g/14	30	0.76	[0.73, 0.79]
Appearance	Attribute	BioCLIP ViT-B/16	30	0.78	[0.75, 0.81]
Appearance	Attribute	Qwen2.5-VL 72B	30	0.52	[0.49, 0.56]
Appearance	Attribute	Qwen2-VL 7B	30	0.51	[0.47, 0.55]
Appearance	Attribute	CLIP ViT-B/16	30	0.81	[0.78, 0.84]
Appearance	Attribute	CLIP ViT-L/14	30	0.76	[0.72, 0.79]
Appearance	Attribute	CLIP ViT-L/14 (336px)	30	0.82	[0.79, 0.85]
Appearance	Attribute	SigLIP ViT-B/16	30	0.82	[0.79, 0.85]
Appearance	Attribute	SigLIP ViT-B/16 (256px)	30	0.81	[0.78, 0.84]
Appearance	Attribute	SigLIP ViT-B/16 (384px)	30	0.80	[0.77, 0.83]

Table 11: All results for 30 training samples for ‘Species’ and ‘Attribute’ tasks.

Task Cluster	Task Subcluster	Model	Train	Mean	Confidence Interval
Appearance	Health	Gemini Flash 2.0	30	0.59	[0.56, 0.63]
Appearance	Health	Gemini Flash 1.5 8B	30	0.54	[0.50, 0.57]
Appearance	Health	SigLIP ViT-B/16 (512px)	30	0.85	[0.83, 0.87]
Appearance	Health	SigLIP ViT-L/16 (256px)	30	0.84	[0.82, 0.86]
Appearance	Health	SigLIP ViT-L/16 (384px)	30	0.87	[0.86, 0.89]
Appearance	Health	SigLIP ViT-SO400M/14	30	0.87	[0.85, 0.89]
Appearance	Health	SigLIP ViT-SO400M/14 (384px)	30	0.89	[0.87, 0.91]
Appearance	Health	DINOv2 ViT-B/14	30	0.83	[0.81, 0.85]
Appearance	Health	DINOv2 ViT-L/14	30	0.83	[0.81, 0.86]
Appearance	Health	DINOv2 ViT-S/14	30	0.82	[0.79, 0.84]
Appearance	Health	DINOv2 ViT-g/14	30	0.83	[0.81, 0.86]
Appearance	Health	BioCLIP ViT-B/16	30	0.81	[0.78, 0.83]
Appearance	Health	Qwen2.5-VL 72B	30	0.49	[0.46, 0.53]
Appearance	Health	Qwen2-VL 7B	30	0.49	[0.46, 0.52]
Appearance	Health	CLIP ViT-B/16	30	0.85	[0.83, 0.88]
Appearance	Health	CLIP ViT-L/14	30	0.86	[0.84, 0.88]
Appearance	Health	CLIP ViT-L/14 (336px)	30	0.87	[0.85, 0.89]
Appearance	Health	SigLIP ViT-B/16	30	0.84	[0.82, 0.86]
Appearance	Health	SigLIP ViT-B/16 (256px)	30	0.86	[0.84, 0.88]
Appearance	Health	SigLIP ViT-B/16 (384px)	30	0.86	[0.83, 0.88]
Appearance	Age	Gemini Flash 2.0	30	0.51	[0.48, 0.54]
Appearance	Age	Gemini Flash 1.5 8B	30	0.50	[0.47, 0.52]
Appearance	Age	SigLIP ViT-B/16 (512px)	30	0.77	[0.75, 0.79]
Appearance	Age	SigLIP ViT-L/16 (256px)	30	0.77	[0.75, 0.79]
Appearance	Age	SigLIP ViT-L/16 (384px)	30	0.79	[0.78, 0.80]
Appearance	Age	SigLIP ViT-SO400M/14	30	0.76	[0.74, 0.78]
Appearance	Age	SigLIP ViT-SO400M/14 (384px)	30	0.78	[0.76, 0.80]
Appearance	Age	DINOv2 ViT-B/14	30	0.84	[0.82, 0.86]
Appearance	Age	DINOv2 ViT-L/14	30	0.82	[0.80, 0.84]
Appearance	Age	DINOv2 ViT-S/14	30	0.83	[0.81, 0.85]
Appearance	Age	DINOv2 ViT-g/14	30	0.82	[0.80, 0.84]
Appearance	Age	BioCLIP ViT-B/16	30	0.75	[0.73, 0.77]
Appearance	Age	Qwen2.5-VL 72B	30	0.51	[0.49, 0.54]
Appearance	Age	Qwen2-VL 7B	30	0.50	[0.47, 0.53]
Appearance	Age	CLIP ViT-B/16	30	0.74	[0.72, 0.76]
Appearance	Age	CLIP ViT-L/14	30	0.74	[0.72, 0.76]
Appearance	Age	CLIP ViT-L/14 (336px)	30	0.75	[0.73, 0.77]
Appearance	Age	SigLIP ViT-B/16	30	0.75	[0.72, 0.77]
Appearance	Age	SigLIP ViT-B/16 (256px)	30	0.76	[0.74, 0.78]
Appearance	Age	SigLIP ViT-B/16 (384px)	30	0.78	[0.76, 0.80]

Table 12: All results for 30 training samples for ‘Health’ and ‘Age’ tasks.

Task Cluster	Task Subcluster	Model	Train	Mean	Confidence Interval
Gestalt	-	Gemini Flash 2.0	30	0.36	$[0.32, 0.40]$
Gestalt	-	Gemini Flash 1.5 8B	30	0.50	$[0.46, 0.54]$
Gestalt	-	SigLIP ViT-B/16 (512px)	30	0.84	$[0.82, 0.85]$
Gestalt	-	SigLIP ViT-L/16 (256px)	30	0.81	$[0.80, 0.83]$
Gestalt	-	SigLIP ViT-L/16 (384px)	30	0.86	$[0.85, 0.87]$
Gestalt	-	SigLIP ViT-SO400M/14	30	0.82	$[0.81, 0.84]$
Gestalt	-	SigLIP ViT-SO400M/14 (384px)	30	0.85	$[0.84, 0.87]$
Gestalt	-	DINOv2 ViT-B/14	30	0.71	$[0.69, 0.73]$
Gestalt	-	DINOv2 ViT-L/14	30	0.60	$[0.58, 0.62]$
Gestalt	-	DINOv2 ViT-S/14	30	0.68	$[0.66, 0.70]$
Gestalt	-	DINOv2 ViT-g/14	30	0.70	$[0.68, 0.72]$
Gestalt	-	BioCLIP ViT-B/16	30	0.65	$[0.63, 0.67]$
Gestalt	-	Qwen2.5-VL 72B	30	0.49	$[0.45, 0.53]$
Gestalt	-	Qwen2-VL 7B	30	0.51	$[0.47, 0.56]$
Gestalt	-	CLIP ViT-B/16	30	0.82	$[0.80, 0.83]$
Gestalt	-	CLIP ViT-L/14	30	0.84	$[0.82, 0.86]$
Gestalt	-	CLIP ViT-L/14 (336px)	30	0.82	$[0.81, 0.84]$
Gestalt	-	SigLIP ViT-B/16	30	0.82	$[0.81, 0.84]$
Gestalt	-	SigLIP ViT-B/16 (256px)	30	0.82	$[0.80, 0.83]$
Gestalt	-	SigLIP ViT-B/16 (384px)	30	0.85	$[0.84, 0.87]$
Context	-	Gemini Flash 2.0	30	0.55	$[0.51, 0.58]$
Context	-	Gemini Flash 1.5 8B	30	0.49	$[0.46, 0.53]$
Context	-	SigLIP ViT-B/16 (512px)	30	0.88	$[0.85, 0.90]$
Context	-	SigLIP ViT-L/16 (256px)	30	0.87	$[0.85, 0.89]$
Context	-	SigLIP ViT-L/16 (384px)	30	0.89	$[0.87, 0.90]$
Context	-	SigLIP ViT-SO400M/14	30	0.90	$[0.87, 0.91]$
Context	-	SigLIP ViT-SO400M/14 (384px)	30	0.91	$[0.89, 0.93]$
Context	-	DINOv2 ViT-B/14	30	0.79	$[0.76, 0.82]$
Context	-	DINOv2 ViT-L/14	30	0.73	$[0.70, 0.76]$
Context	-	DINOv2 ViT-S/14	30	0.74	$[0.71, 0.77]$
Context	-	DINOv2 ViT-g/14	30	0.74	$[0.71, 0.77]$
Context	-	BioCLIP ViT-B/16	30	0.69	$[0.66, 0.73]$
Context	-	Qwen2.5-VL 72B	30	0.50	$[0.47, 0.54]$
Context	-	Qwen2-VL 7B	30	0.48	$[0.45, 0.52]$
Context	-	CLIP ViT-B/16	30	0.82	$[0.79, 0.84]$
Context	-	CLIP ViT-L/14	30	0.84	$[0.81, 0.86]$
Context	-	CLIP ViT-L/14 (336px)	30	0.87	$[0.85, 0.89]$
Context	-	SigLIP ViT-B/16	30	0.84	$[0.82, 0.87]$
Context	-	SigLIP ViT-B/16 (256px)	30	0.84	$[0.82, 0.86]$
Context	-	SigLIP ViT-B/16 (384px)	30	0.87	$[0.85, 0.89]$

Table 13: All results for 30 training samples for ‘Gestalt’ and ‘Context’ tasks.

Task Cluster	Task Subcluster	Model	Train	Mean	Confidence Interval
Counting	-	Gemini Flash 2.0	30	0.49	$[0.43, 0.56]$
Counting	-	Gemini Flash 1.5 8B	30	0.48	$[0.41, 0.55]$
Counting	-	SigLIP ViT-B/16 (512px)	30	0.64	$[0.56, 0.70]$
Counting	-	SigLIP ViT-L/16 (256px)	30	0.67	$[0.60, 0.73]$
Counting	-	SigLIP ViT-L/16 (384px)	30	0.62	$[0.57, 0.67]$
Counting	-	SigLIP ViT-SO400M/14	30	0.64	$[0.57, 0.70]$
Counting	-	SigLIP ViT-SO400M/14 (384px)	30	0.61	$[0.55, 0.69]$
Counting	-	DINOv2 ViT-B/14	30	0.69	$[0.64, 0.74]$
Counting	-	DINOv2 ViT-L/14	30	0.67	$[0.62, 0.71]$
Counting	-	DINOv2 ViT-S/14	30	0.62	$[0.57, 0.67]$
Counting	-	DINOv2 ViT-g/14	30	0.64	$[0.59, 0.69]$
Counting	-	BioCLIP ViT-B/16	30	0.58	$[0.54, 0.63]$
Counting	-	Qwen2.5-VL 72B	30	0.47	$[0.41, 0.55]$
Counting	-	Qwen2-VL 7B	30	0.48	$[0.41, 0.56]$
Counting	-	CLIP ViT-B/16	30	0.63	$[0.59, 0.68]$
Counting	-	CLIP ViT-L/14	30	0.64	$[0.60, 0.69]$
Counting	-	CLIP ViT-L/14 (336px)	30	0.67	$[0.62, 0.71]$
Counting	-	SigLIP ViT-B/16	30	0.62	$[0.56, 0.68]$
Counting	-	SigLIP ViT-B/16 (256px)	30	0.61	$[0.55, 0.68]$
Counting	-	SigLIP ViT-B/16 (384px)	30	0.64	$[0.57, 0.70]$
Behavior	-	Gemini Flash 2.0	30	0.52	$[0.50, 0.55]$
Behavior	-	Gemini Flash 1.5 8B	30	0.49	$[0.46, 0.51]$
Behavior	-	SigLIP ViT-B/16 (512px)	30	0.79	$[0.77, 0.82]$
Behavior	-	SigLIP ViT-L/16 (256px)	30	0.78	$[0.76, 0.80]$
Behavior	-	SigLIP ViT-L/16 (384px)	30	0.83	$[0.82, 0.84]$
Behavior	-	SigLIP ViT-SO400M/14	30	0.79	$[0.77, 0.81]$
Behavior	-	SigLIP ViT-SO400M/14 (384px)	30	0.85	$[0.83, 0.86]$
Behavior	-	DINOv2 ViT-B/14	30	0.66	$[0.64, 0.68]$
Behavior	-	DINOv2 ViT-L/14	30	0.63	$[0.62, 0.65]$
Behavior	-	DINOv2 ViT-S/14	30	0.68	$[0.66, 0.70]$
Behavior	-	DINOv2 ViT-g/14	30	0.66	$[0.64, 0.67]$
Behavior	-	BioCLIP ViT-B/16	30	0.65	$[0.63, 0.67]$
Behavior	-	Qwen2.5-VL 72B	30	0.49	$[0.46, 0.52]$
Behavior	-	Qwen2-VL 7B	30	0.53	$[0.50, 0.56]$
Behavior	-	CLIP ViT-B/16	30	0.68	$[0.66, 0.70]$
Behavior	-	CLIP ViT-L/14	30	0.77	$[0.75, 0.78]$
Behavior	-	CLIP ViT-L/14 (336px)	30	0.81	$[0.79, 0.82]$
Behavior	-	SigLIP ViT-B/16	30	0.73	$[0.71, 0.75]$
Behavior	-	SigLIP ViT-B/16 (256px)	30	0.76	$[0.73, 0.78]$
Behavior	-	SigLIP ViT-B/16 (384px)	30	0.78	$[0.76, 0.80]$

Table 14: All results for 30 training samples for ‘Counting’ and ‘Behavior’ tasks.

Task Cluster	Task Subcluster	Model	Train	Mean	Confidence Interval
Appearance	Species	Gemini Flash 2.0	100	0.48	$[0.48, 0.49]$
Appearance	Species	Gemini Flash 1.5 8B	100	0.50	$[0.49, 0.51]$
Appearance	Species	SigLIP ViT-B/16 (512px)	100	0.80	$[0.79, 0.80]$
Appearance	Species	SigLIP ViT-L/16 (256px)	100	0.79	$[0.79, 0.80]$
Appearance	Species	SigLIP ViT-L/16 (384px)	100	0.81	$[0.81, 0.82]$
Appearance	Species	SigLIP ViT-SO400M/14	100	0.81	$[0.80, 0.81]$
Appearance	Species	SigLIP ViT-SO400M/14 (384px)	100	0.82	$[0.82, 0.83]$
Appearance	Species	DINOv2 ViT-B/14	100	0.84	$[0.83, 0.84]$
Appearance	Species	DINOv2 ViT-L/14	100	0.84	$[0.83, 0.84]$
Appearance	Species	DINOv2 ViT-S/14	100	0.81	$[0.80, 0.81]$
Appearance	Species	DINOv2 ViT-g/14	100	0.86	$[0.86, 0.87]$
Appearance	Species	BioCLIP ViT-B/16	100	0.83	$[0.82, 0.83]$
Appearance	Species	Qwen2.5-VL 72B	100	0.49	$[0.48, 0.50]$
Appearance	Species	Qwen2-VL 7B	100	0.50	$[0.50, 0.51]$
Appearance	Species	CLIP ViT-B/16	100	0.76	$[0.75, 0.76]$
Appearance	Species	CLIP ViT-L/14	100	0.78	$[0.78, 0.79]$
Appearance	Species	CLIP ViT-L/14 (336px)	100	0.79	$[0.78, 0.79]$
Appearance	Species	SigLIP ViT-B/16	100	0.78	$[0.77, 0.78]$
Appearance	Species	SigLIP ViT-B/16 (256px)	100	0.78	$[0.78, 0.79]$
Appearance	Species	SigLIP ViT-B/16 (384px)	100	0.79	$[0.79, 0.80]$
Appearance	Attribute	Gemini Flash 2.0	100	0.52	$[0.49, 0.56]$
Appearance	Attribute	Gemini Flash 1.5 8B	100	0.50	$[0.46, 0.54]$
Appearance	Attribute	SigLIP ViT-B/16 (512px)	100	0.90	$[0.89, 0.92]$
Appearance	Attribute	SigLIP ViT-L/16 (256px)	100	0.89	$[0.88, 0.91]$
Appearance	Attribute	SigLIP ViT-L/16 (384px)	100	0.89	$[0.88, 0.90]$
Appearance	Attribute	SigLIP ViT-SO400M/14	100	0.89	$[0.87, 0.90]$
Appearance	Attribute	SigLIP ViT-SO400M/14 (384px)	100	0.90	$[0.88, 0.91]$
Appearance	Attribute	DINOv2 ViT-B/14	100	0.86	$[0.85, 0.88]$
Appearance	Attribute	DINOv2 ViT-L/14	100	0.82	$[0.80, 0.84]$
Appearance	Attribute	DINOv2 ViT-S/14	100	0.88	$[0.86, 0.89]$
Appearance	Attribute	DINOv2 ViT-g/14	100	0.85	$[0.83, 0.87]$
Appearance	Attribute	BioCLIP ViT-B/16	100	0.86	$[0.85, 0.88]$
Appearance	Attribute	Qwen2.5-VL 72B	100	0.49	$[0.45, 0.53]$
Appearance	Attribute	Qwen2-VL 7B	100	0.49	$[0.47, 0.51]$
Appearance	Attribute	CLIP ViT-B/16	100	0.88	$[0.86, 0.89]$
Appearance	Attribute	CLIP ViT-L/14	100	0.87	$[0.85, 0.89]$
Appearance	Attribute	CLIP ViT-L/14 (336px)	100	0.91	$[0.89, 0.92]$
Appearance	Attribute	SigLIP ViT-B/16	100	0.86	$[0.84, 0.87]$
Appearance	Attribute	SigLIP ViT-B/16 (256px)	100	0.85	$[0.84, 0.87]$
Appearance	Attribute	SigLIP ViT-B/16 (384px)	100	0.88	$[0.87, 0.89]$

Table 15: All results for 100 training samples for ‘Species’ and ‘Attribute’ tasks.

Task Cluster	Task Subcluster	Model	Train	Mean	Confidence Interval
Appearance	Health	Gemini Flash 2.0	100	0.46	[0.43, 0.50]
Appearance	Health	Gemini Flash 1.5 8B	100	0.54	[0.51, 0.58]
Appearance	Health	SigLIP ViT-B/16 (512px)	100	0.88	[0.87, 0.89]
Appearance	Health	SigLIP ViT-L/16 (256px)	100	0.88	[0.87, 0.89]
Appearance	Health	SigLIP ViT-L/16 (384px)	100	0.90	[0.89, 0.90]
Appearance	Health	SigLIP ViT-SO400M/14	100	0.89	[0.87, 0.90]
Appearance	Health	SigLIP ViT-SO400M/14 (384px)	100	0.90	[0.89, 0.91]
Appearance	Health	DINOv2 ViT-B/14	100	0.90	[0.89, 0.91]
Appearance	Health	DINOv2 ViT-L/14	100	0.88	[0.87, 0.89]
Appearance	Health	DINOv2 ViT-S/14	100	0.88	[0.87, 0.89]
Appearance	Health	DINOv2 ViT-g/14	100	0.89	[0.88, 0.90]
Appearance	Health	BioCLIP ViT-B/16	100	0.85	[0.84, 0.87]
Appearance	Health	Qwen2.5-VL 72B	100	0.49	[0.46, 0.52]
Appearance	Health	Qwen2-VL 7B	100	0.50	[0.48, 0.52]
Appearance	Health	CLIP ViT-B/16	100	0.86	[0.85, 0.88]
Appearance	Health	CLIP ViT-L/14	100	0.88	[0.87, 0.90]
Appearance	Health	CLIP ViT-L/14 (336px)	100	0.92	[0.90, 0.93]
Appearance	Health	SigLIP ViT-B/16	100	0.86	[0.85, 0.88]
Appearance	Health	SigLIP ViT-B/16 (256px)	100	0.87	[0.86, 0.88]
Appearance	Health	SigLIP ViT-B/16 (384px)	100	0.88	[0.87, 0.90]
Appearance	Age	Gemini Flash 2.0	100	0.48	[0.45, 0.50]
Appearance	Age	Gemini Flash 1.5 8B	100	0.51	[0.48, 0.53]
Appearance	Age	SigLIP ViT-B/16 (512px)	100	0.89	[0.88, 0.90]
Appearance	Age	SigLIP ViT-L/16 (256px)	100	0.86	[0.85, 0.87]
Appearance	Age	SigLIP ViT-L/16 (384px)	100	0.87	[0.86, 0.88]
Appearance	Age	SigLIP ViT-SO400M/14	100	0.85	[0.84, 0.86]
Appearance	Age	SigLIP ViT-SO400M/14 (384px)	100	0.89	[0.88, 0.90]
Appearance	Age	DINOv2 ViT-B/14	100	0.92	[0.91, 0.93]
Appearance	Age	DINOv2 ViT-L/14	100	0.93	[0.92, 0.94]
Appearance	Age	DINOv2 ViT-S/14	100	0.92	[0.92, 0.93]
Appearance	Age	DINOv2 ViT-g/14	100	0.93	[0.93, 0.94]
Appearance	Age	BioCLIP ViT-B/16	100	0.87	[0.86, 0.88]
Appearance	Age	Qwen2.5-VL 72B	100	0.48	[0.46, 0.51]
Appearance	Age	Qwen2-VL 7B	100	0.49	[0.48, 0.51]
Appearance	Age	CLIP ViT-B/16	100	0.83	[0.82, 0.85]
Appearance	Age	CLIP ViT-L/14	100	0.84	[0.83, 0.85]
Appearance	Age	CLIP ViT-L/14 (336px)	100	0.87	[0.86, 0.88]
Appearance	Age	SigLIP ViT-B/16	100	0.85	[0.84, 0.86]
Appearance	Age	SigLIP ViT-B/16 (256px)	100	0.85	[0.84, 0.87]
Appearance	Age	SigLIP ViT-B/16 (384px)	100	0.88	[0.87, 0.89]

Table 16: All results for 100 training samples for ‘Health’ and ‘Age’ tasks.

Task Cluster	Task Subcluster	Model	Train	Mean	Confidence Interval
Gestalt	-	Gemini Flash 2.0	100	0.45	$[0.41, 0.49]$
Gestalt	-	Gemini Flash 1.5 8B	100	0.50	$[0.46, 0.54]$
Gestalt	-	SigLIP ViT-B/16 (512px)	100	0.88	$[0.87, 0.89]$
Gestalt	-	SigLIP ViT-L/16 (256px)	100	0.87	$[0.86, 0.89]$
Gestalt	-	SigLIP ViT-L/16 (384px)	100	0.89	$[0.89, 0.90]$
Gestalt	-	SigLIP ViT-SO400M/14	100	0.87	$[0.86, 0.89]$
Gestalt	-	SigLIP ViT-SO400M/14 (384px)	100	0.89	$[0.88, 0.90]$
Gestalt	-	DINOv2 ViT-B/14	100	0.81	$[0.80, 0.83]$
Gestalt	-	DINOv2 ViT-L/14	100	0.77	$[0.75, 0.79]$
Gestalt	-	DINOv2 ViT-S/14	100	0.84	$[0.83, 0.86]$
Gestalt	-	DINOv2 ViT-g/14	100	0.74	$[0.72, 0.76]$
Gestalt	-	BioCLIP ViT-B/16	100	0.77	$[0.75, 0.79]$
Gestalt	-	Qwen2.5-VL 72B	100	0.53	$[0.49, 0.57]$
Gestalt	-	Qwen2-VL 7B	100	0.54	$[0.51, 0.56]$
Gestalt	-	CLIP ViT-B/16	100	0.88	$[0.87, 0.90]$
Gestalt	-	CLIP ViT-L/14	100	0.88	$[0.86, 0.89]$
Gestalt	-	CLIP ViT-L/14 (336px)	100	0.90	$[0.89, 0.91]$
Gestalt	-	SigLIP ViT-B/16	100	0.89	$[0.88, 0.90]$
Gestalt	-	SigLIP ViT-B/16 (256px)	100	0.88	$[0.86, 0.89]$
Gestalt	-	SigLIP ViT-B/16 (384px)	100	0.89	$[0.88, 0.90]$
Context	-	Gemini Flash 2.0	100	0.51	$[0.47, 0.54]$
Context	-	Gemini Flash 1.5 8B	100	0.51	$[0.47, 0.54]$
Context	-	SigLIP ViT-B/16 (512px)	100	0.93	$[0.92, 0.94]$
Context	-	SigLIP ViT-L/16 (256px)	100	0.93	$[0.92, 0.94]$
Context	-	SigLIP ViT-L/16 (384px)	100	0.93	$[0.92, 0.94]$
Context	-	SigLIP ViT-SO400M/14	100	0.94	$[0.93, 0.95]$
Context	-	SigLIP ViT-SO400M/14 (384px)	100	0.95	$[0.94, 0.96]$
Context	-	DINOv2 ViT-B/14	100	0.80	$[0.78, 0.81]$
Context	-	DINOv2 ViT-L/14	100	0.79	$[0.78, 0.81]$
Context	-	DINOv2 ViT-S/14	100	0.83	$[0.81, 0.84]$
Context	-	DINOv2 ViT-g/14	100	0.79	$[0.78, 0.81]$
Context	-	BioCLIP ViT-B/16	100	0.76	$[0.74, 0.78]$
Context	-	Qwen2.5-VL 72B	100	0.48	$[0.45, 0.52]$
Context	-	Qwen2-VL 7B	100	0.49	$[0.47, 0.51]$
Context	-	CLIP ViT-B/16	100	0.88	$[0.87, 0.89]$
Context	-	CLIP ViT-L/14	100	0.92	$[0.91, 0.93]$
Context	-	CLIP ViT-L/14 (336px)	100	0.93	$[0.92, 0.94]$
Context	-	SigLIP ViT-B/16	100	0.91	$[0.90, 0.92]$
Context	-	SigLIP ViT-B/16 (256px)	100	0.92	$[0.91, 0.93]$
Context	-	SigLIP ViT-B/16 (384px)	100	0.93	$[0.92, 0.94]$

Table 17: All results for 100 training samples for ‘Gestalt’ and ‘Context’ tasks.

Task Cluster	Task Subcluster	Model	Train	Mean	Confidence Interval
Counting	-	Gemini Flash 2.0	100	0.52	$[0.45, 0.58]$
Counting	-	Gemini Flash 1.5 8B	100	0.45	$[0.38, 0.52]$
Counting	-	SigLIP ViT-B/16 (512px)	100	0.80	$[0.77, 0.83]$
Counting	-	SigLIP ViT-L/16 (256px)	100	0.80	$[0.76, 0.83]$
Counting	-	SigLIP ViT-L/16 (384px)	100	0.85	$[0.83, 0.88]$
Counting	-	SigLIP ViT-SO400M/14	100	0.81	$[0.78, 0.84]$
Counting	-	SigLIP ViT-SO400M/14 (384px)	100	0.84	$[0.81, 0.87]$
Counting	-	DINOv2 ViT-B/14	100	0.73	$[0.70, 0.75]$
Counting	-	DINOv2 ViT-L/14	100	0.77	$[0.74, 0.79]$
Counting	-	DINOv2 ViT-S/14	100	0.69	$[0.66, 0.71]$
Counting	-	DINOv2 ViT-g/14	100	0.72	$[0.70, 0.75]$
Counting	-	BioCLIP ViT-B/16	100	0.60	$[0.58, 0.63]$
Counting	-	Qwen2.5-VL 72B	100	0.51	$[0.43, 0.57]$
Counting	-	Qwen2-VL 7B	100	0.51	$[0.47, 0.55]$
Counting	-	CLIP ViT-B/16	100	0.61	$[0.59, 0.64]$
Counting	-	CLIP ViT-L/14	100	0.72	$[0.70, 0.75]$
Counting	-	CLIP ViT-L/14 (336px)	100	0.75	$[0.73, 0.77]$
Counting	-	SigLIP ViT-B/16	100	0.76	$[0.72, 0.79]$
Counting	-	SigLIP ViT-B/16 (256px)	100	0.77	$[0.73, 0.80]$
Counting	-	SigLIP ViT-B/16 (384px)	100	0.73	$[0.69, 0.77]$
Behavior	-	Gemini Flash 2.0	100	0.46	$[0.43, 0.48]$
Behavior	-	Gemini Flash 1.5 8B	100	0.52	$[0.49, 0.55]$
Behavior	-	SigLIP ViT-B/16 (512px)	100	0.86	$[0.85, 0.87]$
Behavior	-	SigLIP ViT-L/16 (256px)	100	0.86	$[0.85, 0.87]$
Behavior	-	SigLIP ViT-L/16 (384px)	100	0.89	$[0.88, 0.90]$
Behavior	-	SigLIP ViT-SO400M/14	100	0.86	$[0.85, 0.87]$
Behavior	-	SigLIP ViT-SO400M/14 (384px)	100	0.88	$[0.87, 0.89]$
Behavior	-	DINOv2 ViT-B/14	100	0.78	$[0.77, 0.79]$
Behavior	-	DINOv2 ViT-L/14	100	0.72	$[0.71, 0.73]$
Behavior	-	DINOv2 ViT-S/14	100	0.76	$[0.75, 0.77]$
Behavior	-	DINOv2 ViT-g/14	100	0.74	$[0.74, 0.75]$
Behavior	-	BioCLIP ViT-B/16	100	0.74	$[0.74, 0.75]$
Behavior	-	Qwen2.5-VL 72B	100	0.49	$[0.47, 0.52]$
Behavior	-	Qwen2-VL 7B	100	0.52	$[0.50, 0.53]$
Behavior	-	CLIP ViT-B/16	100	0.77	$[0.76, 0.78]$
Behavior	-	CLIP ViT-L/14	100	0.86	$[0.85, 0.86]$
Behavior	-	CLIP ViT-L/14 (336px)	100	0.86	$[0.86, 0.87]$
Behavior	-	SigLIP ViT-B/16	100	0.80	$[0.79, 0.81]$
Behavior	-	SigLIP ViT-B/16 (256px)	100	0.82	$[0.81, 0.83]$
Behavior	-	SigLIP ViT-B/16 (384px)	100	0.85	$[0.84, 0.86]$

Table 18: All results for 100 training samples for ‘Counting’ and ‘Behavior’ tasks.