PrimateFace: A Resource for Generalizable Cross-Species Facial Analysis

F. Parodi[1] **J.K. Matelsky**[1,2] **A.P. Lamacchia**[1] **M. Segado**[1]

Y. Jiang[1] **A. Regla-Vargas**[1] **L. Sofi**[1] **C. Kimock**[3]

B.M. Waller[3] **M.L. Platt**[1] **K.P. Kording**[1,4]

[1]University of Pennsylvania
[2]Johns Hopkins University Applied Physics Laboratory
[3]Nottingham Trent University
[4]CIFAR

Abstract

The analysis of facial kinematics, a critical class of behavioral biosignals, is fundamental to understanding social cognition. Progress has been hampered by the lack of large-scale, diverse datasets needed to train robust and generalizable models for face detection and facial landmark estimation (FLE). We introduce PrimateFace, a resource designed to overcome this bottleneck. It consists of a large-scale dataset of over 260,000 images spanning more than 60 primate genera, and a suite of pretrained models. Models trained on PrimateFace achieve high cross-species performance, from tarsiers to gorillas, and demonstrate remarkable generalization to human data, validating the benefits of pre-training on taxonomically diverse data. We showcase how PrimateFace serves as an essential front-end for diverse downstream applications, including quantifying social gaze in human infants, enabling multimodal analysis of vocalizations, and powering the data-driven discovery of behavioral repertoires. PrimateFace provides a standardized platform for extracting and analyzing behavioral biosignals, empowering scalable, data-driven studies of behavior.

1 Introduction

Faces are essential conduits for conveying information critical to social animals, with relevance to psychology, neuroscience, evolutionary biology, and conservation. Among the many signals the brain and body produce, facial movements are a particularly high-dimensional and information-rich stream of data. The primary challenge lies in reliably quantifying this signal by transforming raw video into a structured, continuous kinematic biosignal—a process fundamental to decoding the intricate dynamics of social behavior.

While deep learning has driven progress in human facial analysis, powered by massive datasets like WIDERFace (Yang and others [2016]) and COCO-WholeBody (Jin and others [2020]), equivalent tools for non-human primates remain limited (Bala and others [2020], Carugati et al. [2025]). Existing approaches are typically trained on small, taxonomically narrow datasets, leading to specialized models that fail to generalize across the immense morphological heterogeneity of the primate order (Schofield and others [2023]).

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

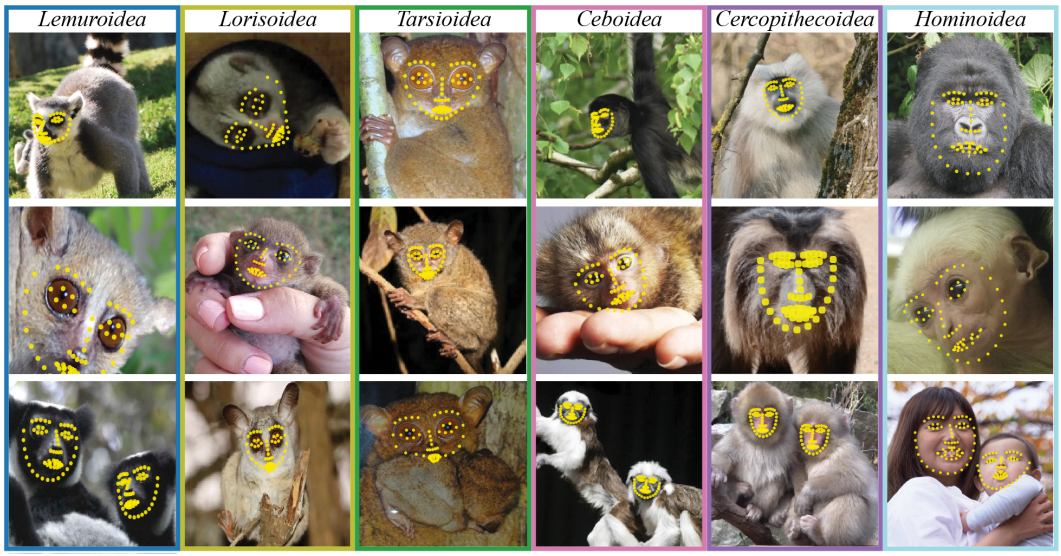

Figure 1: PrimateFace dataset provides a diverse foundation for cross-species analysis.

This diversity, coupled with the limitations of labor-intensive manual coding methods like the Facial Action Coding System (FACS) (Waller et al. [2020], has made large-scale, comparative studies of behavior computationally intractable. A foundational resource—a large-scale, taxonomically diverse pretraining dataset—is required to learn truly generalizable representations.

Here, we introduce PrimateFace, a foundational resource designed to address this challenge and accelerate research in behavioral biosignal analysis. Our contributions are threefold:

1. We constructed the largest and most taxonomically diverse dataset of annotated primate faces, designed for pretraining models that learn generalizable representations.
2. We developed a suite of pretrained models and demonstrate that their powerful generalization properties stem from pre-training on taxonomically diverse data.
3. We showcase the utility of PrimateFace as a front-end for diverse scientific applications, from cross-species behavioral analysis to automated individual recognition.

2 A Cross-Species Dataset for Pretraining

The foundation of our resource is a large-scale, taxonomically diverse dataset curated specifically for pretraining generalizable models, comprising over 260,000 images of more than 60 primate genera. We prioritized taxonomic breadth, spanning all six primate superfamilies, from Lemuroidea to Hominoidea (Figure 1). Exposing models to this diversity is critical for learning the invariant features that define a primate face, forcing the model to move beyond species-specific traits and develop a more fundamental understanding of primate facial structure, which is the key to generalization.

The annotations in PrimateFace are designed to transform unstructured pixel data into structured, analyzable biosignals Figure A.1. Every face is annotated with a bounding box and a standardized 68-point landmark configuration. When applied to video, models trained on this data produce a continuous time-series of landmark coordinates – a high-dimensional kinematic biosignal that serves as the input for downstream analysis.

3 Generalization Properties of Models Trained on PrimateFace

We trained several computer vision models, including models from the OpenMMLab ecosystem, DeepLabCut (Mathis and others [2018]), SLEAP (Pereira and others [2022]) and Ultralytics (e.g., Zhang and others [2025]) and evaluated their generalization properties.

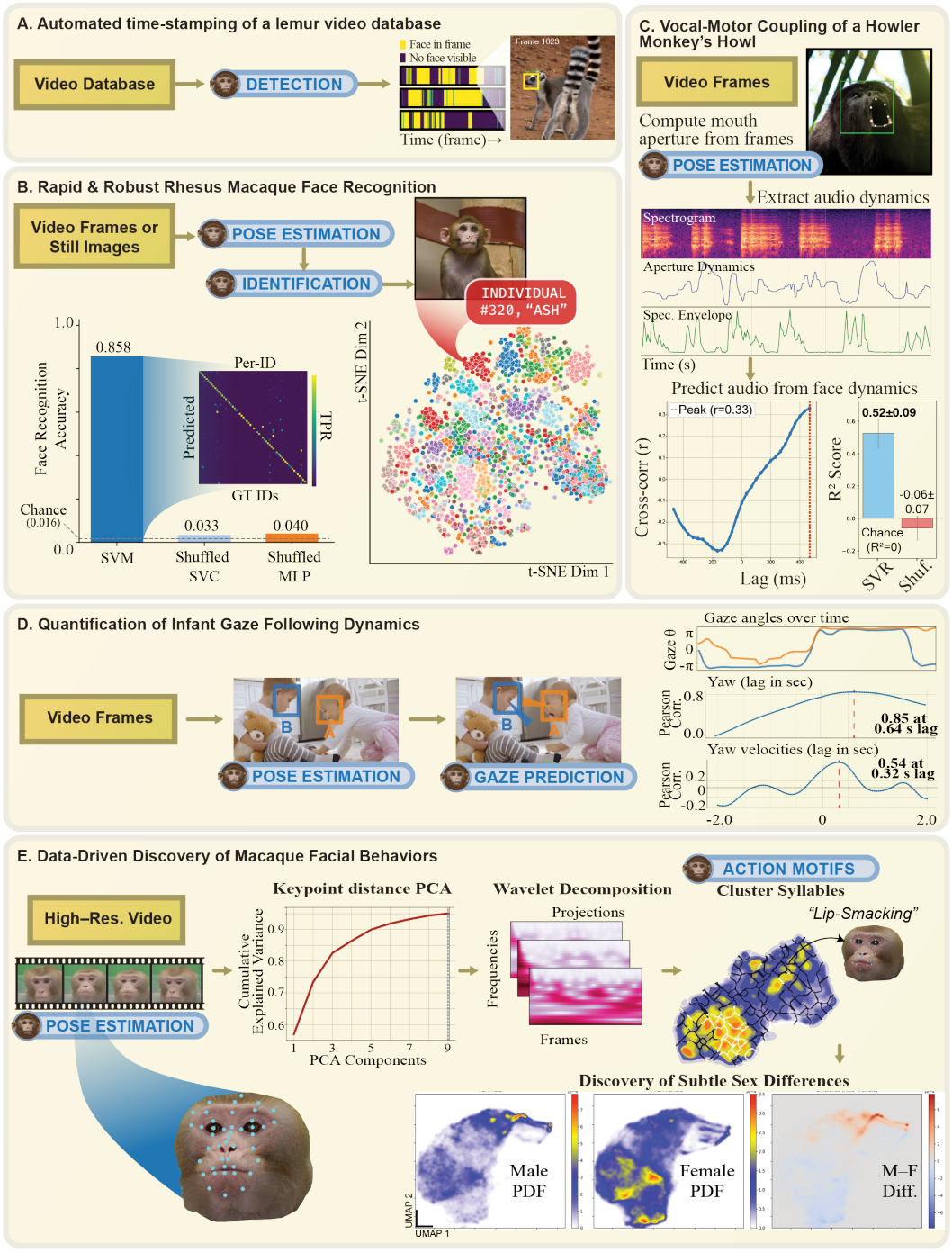

Figure 2: PrimateFace accelerates and enables diverse research applications.

Our experiments confirm that pretraining on the taxonomically diverse PrimateFace dataset yields models with remarkable generalization. When evaluated on the challenging COCO-WholeBody-Face human benchmark in a zero-shot setting, our PrimateFace-trained model achieves a Normalized Mean Error (NME) of 0.061, performing competitively with specialist models trained exclusively on human data (0.053 NME) Figure A.2.

This generalization is notably asymmetric; a model trained only on human data exhibits a substantial performance degradation when evaluated on our diverse primate dataset (0.122 NME vs. 0.029 NME) (Figure A.3). This result illustrates the benefit of our approach: **pretraining on morphologically diverse data yields general representations that transfer broadly, whereas narrow pretraining results in specialized models that fail to generalize.**

4 Downstream Applications Enabled by PrimateFace

To demonstrate the utility of PrimateFace as a foundational resource, we present its application in five distinct biosignal analysis and scientific automation domains. In Figure 2, we showcase these use cases, ranging from A. lemur face tracking in the wild; B. rapid face recognition of macaque monkeys; C. vocal-motor coupling analysis of a Howler Monkey's vocalizations; D. social gaze analysis in human infants; and E. unsupervised identification of facial action motifs. Below, we detail these applications.

4.1 Scientific Automation: Automated Time-Stamping.

Manually logging individual presence in video is a time-consuming bottleneck in observational research. PrimateFace's cross-species detectors automate this critical step. Our pipeline processes video frame-by-frame to effectively compress days of raw footage into concise visualizations of individual visibility over time, enhancing the efficiency of longitudinal monitoring.

4.2 Scientific Automation: Rapid Individual Recognition.

Building individual recognition systems is historically labor-intensive. PrimateFace automates the critical front-end steps of detection and alignment. Using our models, a pipeline to detect, align, and generate embeddings for a classifier was executed on a public dataset of 62 macaques in under an hour, achieving 0.858 top-1 accuracy. This demonstrates how PrimateFace enables researchers to rapidly create accurate ID systems for large cohorts without specialized development.

4.3 Cross-Signal Analysis: Vocal-Motor Coupling of Howler Monkey's Howl.

Understanding vocal communication requires coordinating facial movements and sound. We applied our FLE model to extract a continuous kinematic biosignal of mouth aperture from video of a howling howler monkey. Aligning this with the acoustic biosignal (the spectrogram's temporal envelope) allows for the quantification of precise coupling between mouth motion and vocal output, enabling a more mechanistic understanding of vocal production.

4.4 Cross-Species Generalization: Quantifying Social Gaze in Human Infants.

Analyzing joint attention in developmental psychology is notoriously labor-intensive. Our resource's demonstrated cross-species generalization allows us to apply PrimateFace models "off-the-shelf" to human data. We use our robust face detector to track interacting infants, which then serve as inputs to a specialized gaze estimation model (Ryan and others [2024]). This pipeline enables automated, objective, and scalable quantification of fine-grained behavioral synchrony.

4.5 Cross-Subject, Data-Driven Discovery of Facial Action Motifs.

Traditional ethological approaches rely on pre-defined behavioral categories that may miss subtle patterns. PrimateFace's precise landmark tracking enables 'behavioral syllable' discovery from facial kinematics. Using a pipeline inspired by traditional unsupervised approaches (Berman and others [2014]), we automatically identified over 80 recurrent movement patterns from high-resolution

macaque video, discovering both stereotyped movements (e.g., lip-smacking) and subtle, previously unobserved sex-specific differences in communication repertoires.

5 Discussion

Here, we introduced PrimateFace, a foundational resource for the analysis of facial kinematics as a behavioral biosignal. We have shown that pretraining models on a large-scale, taxonomically diverse dataset is a critical step towards overcoming the generalization failures of previous species-specific approaches. This work provides the first empirical proof of a **'taxonomic scaling' principle** for building foundation models for behavior, where generalization scales with morphological heterogeneity.

The implications of this resource extend across multiple fields, including primatology, anthropology, psychology, and neuroscience. For example, a major challenge is building neural decoders that generalize across individuals, due to high inter-subject variability in neural recordings.

Our ongoing ablation studies, such as leave-one-superfamily-out cross-validation, will further probe the properties of taxonomic scaling. Our model's limitations also define key next steps: as a 2D system, it cannot fully disentangle expression from pose, motivating the development of 3D models. Furthermore, performance correlates with taxonomic density, highlighting the need for continued, targeted data collection for the most morphologically unique genera. Finally, we are exploring modern image generation models to augment PrimateFace with synthetic data.

Finally, although face analysis technologies carry risks of misuse, all human data used in our demonstrations were from licensed stock footage. PrimateFace includes lightweight models that can be run on consumer-grade devices, making these tools broadly accessible and suitable for on-device processing. Ultimately, we hope PrimateFace will empower a new generation of scalable, data-driven studies into the intricate links between brain, body, and behavior.

Acknowledgments and Disclosure of Funding

We would like to thank the labs of Michael L. Platt and Konrad P. Kording for their valuable advice and support. We also acknowledge the support of our funding agencies and institutions.

References

Praneet C. Bala and others. Automated markerless pose estimation in freely moving macaques with OpenMonkeyStudio. *Nature communications*, 11:1, 2020.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66, May 2018. doi: 10.1109/FG.2018.00019. URL https://ieeexplore.ieee.org/abstract/document/8373812.

Gordon J. Berman and others. Mapping the stereotyped behaviour of freely moving fruit flies. *Journal of The Royal Society Interface*, 11:99, 2014.

Filippo Carugati, Olivier Friard, Elisa Protopapa, Camilla Mancassola, Emanuela Rabajoli, Chiara De Gregorio, Daria Valente, Valeria Ferrario, Walter Cristiano, Teresa Raimondi, Valeria Torti, Brice Lefaux, Longondraza Miaretsoa, Cristina Giacoma, and Marco Gamba. Discrimination between the facial gestures of vocalising and non-vocalising lemurs and small apes using deep learning. *Ecological Informatics*, 85:102847, March 2025. ISSN 1574-9541. doi: 10.1016/j.ecoinf.2024.102847. URL https://www.sciencedirect.com/science/article/pii/S1574954124003893.

Sheng Jin and others. Whole-body human pose estimation in the wild. In *European Conference on Computer Vision. Cham*. Springer International Publishing, 2020.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, volume 15105, pages 38–55. Springer Nature Switzerland, Cham, 2025. ISBN 978-3-031-72969-0 978-3-031-72970-6. doi: 10.1007/978-3-031-72970-6_3. URL https://link.springer.com/10.1007/978-3-031-72970-6_3. Series Title: Lecture Notes in Computer Science.

Alexander Mathis and others. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289, 2018.

Maxime Oquab and others. Dinov2: Learning robust visual features without supervision. _eprint: 2304.07193, 2023.

Talmo D. Pereira and others. *SLEAP: A deep learning system for multi-animal pose tracking*. 1-10, Nature methods, 2022.

Fiona Ryan and others. Gaze-LLE: Gaze Target Estimation via Large-Scale Learned Encoders. _eprint: 2412.09586, 2024.

Daniel P. Schofield and others. Automated face recognition using deep neural networks produces robust primate social networks and sociality measures. *Methods in Ecology and Evolution*, 14(8): 1937–1951, 2023.

B.M. Waller, E. Julle-Daniere, and J. Micheletta. Measuring the evolution of facial 'expression' using multi-species FACS. *Neuroscience & Biobehavioral Reviews*, 113:1–11, June 2020. ISSN 01497634. doi: 10.1016/j.neubiorev.2020.02.031. URL https://linkinghub.elsevier.com/retrieve/pii/S0149763419302404.

Shuo Yang and others. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

Jia-Jin Zhang and others. A deep learning lightweight model for real-time captive macaque facial recognition based on an improved YOLOX model. *Zoological Research*, 46:2, 2025.

A Technical Appendices and Supplementary Material

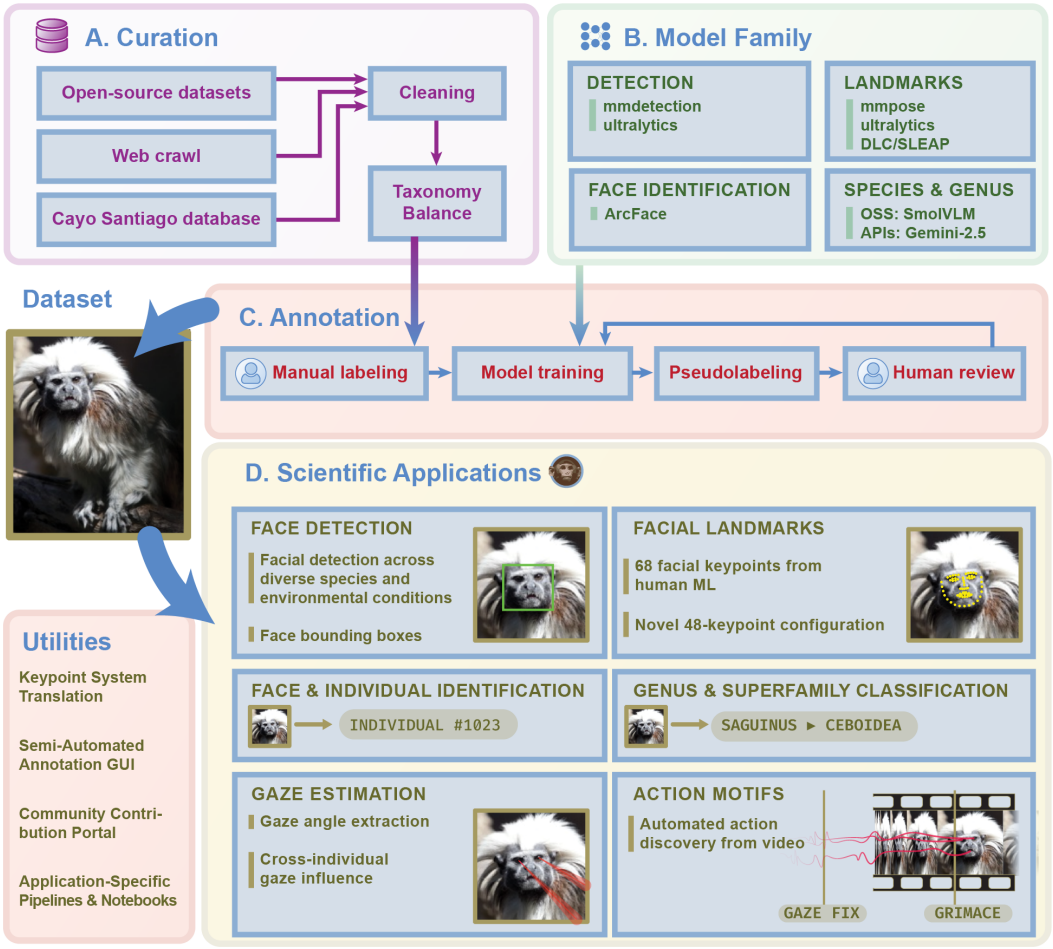

Figure A.1: **The PrimateFace Ecosystem.** An overview of our integrated workflow for building a foundational resource. The process unifies (A) large-scale data curation, (B) development of models using multiple open-source frameworks, and (C) a scalable, semi-automated annotation pipeline. This iterative loop enables the creation of (D) diverse downstream scientific applications.

A.1 PrimateFace Dataset Significance

Prior to PrimateFace, no dataset captured the vast morphological and expressive diversity of primates. The PrimateFace dataset encompasses over 260,000 images: an extensive aggregation of 200,000+ images sourced and re-annotated from diverse public datasets, augmented by a novel collection of 60,000 images sampled for taxonomic balance across 60+ primate genera. This effort uniquely extends coverage across the primate order, including previously underrepresented groups such as tarsiers, alongside robust representation of the superfamilies: Lemuroidea, Lorisoidea, Tarsioidea, Ceboidea, Cercopithecoidea, and Hominoidea. PrimateFace captures this taxonomic breadth across a wide array of ecologically-relevant conditions, including images from wild, captive, and laboratory environments, depicting a spectrum of naturalistic social interactions, developmental stages from infancy to adulthood, and diverse facial expressions, all under varied lighting and image quality (Figure 1). This combination of scale, taxonomic balancing, and contextual richness establishes PrimateFace as a uniquely versatile dataset, poised to unlock frontiers in the automated quantitative study of primate facial behavior.

Achieving comprehensive taxonomic coverage across 60+ genera presented a critical sampling challenge. Traditional random sampling approaches are prohibitively inefficient for taxonomically

imbalanced datasets, often requiring full annotation before capturing sufficient diversity within rare taxa like tarsiers or aye-ayes. To address this, we leveraged DINOv2 (Oquab and others [2023]), a self-supervised vision transformer that captures semantic visual features, to identify and prioritize representationally diverse images across the primate order. Images grouped by DINOv2 image feature embeddings frequently corresponded to distinct genera or species, validating that the approach captured meaningful visual cross-primate diversity. This allowed strategic sampling that maximized taxonomic and morphological diversity with minimal annotation overhead.

Empirical validation demonstrated advantages over random sampling in annotation-constrained scenarios. Across multiple labeling budgets (500–8,000 images), Cascade-RCNN face detectors trained on DINOv2-sampled subsets achieved 2-4 point improvements in mAP@[.50:.95] over random sampling, with peak advantage around 500 training images (Figure A.2). This finding informed our dataset construction strategy: we targeted 500-850 images per genus to maximize model performance while maintaining taxonomic breadth. While both approaches converged at larger dataset sizes, DINOv2 sampling provided essential efficiency gains required for taxonomically diverse, resource-constrained primate research.

This data-driven approach, combined with semi-automated annotation workflows, enabled rapid labeling of face bounding boxes and facial landmarks. Every primate face received a tight bounding box annotation around the face (typically excluding ears), accompanied by both our novel 48-facial-keypoint scheme optimized for cross-species analysis and the standard 68-keypoint format for interoperability with existing pipelines. However, designing landmark schemes that capture morphological diversity from prosimian rostra to great ape facial profiles required careful consideration of existing annotation conventions.

Facial landmark conventions have evolved from rudimentary alignment schemes to increasingly sophisticated anatomical mappings. Early computer vision approaches relied on sparse 5-keypoint configurations sufficient for basic face alignment tasks (Baltrusaitis et al. [2018]), while human facial analysis matured around the 68-keypoint standard popularized by datasets like COCO-WholeBody (Jin and others [2020]) (Figure A.2, left to right). Recognizing the unique challenges of non-human primate morphology, prior work has developed NHP-specific landmark schemes, most notably DeepLabCut's 55-keypoint MacaqueFace model in their Model Zoo (Mathis and others [2018]) (Figure A.2, second from the right) – a pioneering effort that enabled essential pseudo-labeling for our annotation pipeline. However, no existing convention adequately captures the morphological heterogeneity spanning the entire primate order, from pronounced baboon muzzles to flat-nosed capuchins.

PrimateFace addresses this gap through rich, dual-layered annotations designed to support both specialized primatological inquiry and broader computer vision applications. The 48-keypoint configuration (Figure A.2 far right), derived by combining the most robust landmarks from both DeepLabCut's MacaqueFace model and the human 68-keypoint standard, strategically emphasizes anatomical landmarks that meet three key criteria: 1) consistent visibility across species (excluding neck nape landmarks often occluded by posture); 2) anatomical precision (omitting COCO-68 chin contours that lack clear anatomical correspondence even in human faces); and 3) universal applicability, namely the exclusion of discrete pupil landmarks since many primate species lack clearly visible pupils. This configuration retains valuable detailed landmarks from the COCO-68 standard, particularly the precise mouth contour and nose bridge annotations essential for facial expression analysis, while ensuring reliable cross-species applicability. Researchers can leverage subsets of these landmarks for region-specific analyses, such as focusing on periocular keypoints for eye movement studies or mouth contours for vocal expression analysis (see Scientific Application 3). The parallel 68-keypoint annotations maintain interoperability with existing human facial analysis pipelines, creating a unified foundation for developing future cross-species models for primate behavior analysis. All annotations are exportable to common movement analysis frameworks such as SLEAP (Pereira and others [2022]), DeepLabCut (Mathis and others [2018]), OpenMMLab, or Ultralytics. Researchers are encouraged to contribute data at our community contribution portal (github.com/KordingLab/PrimateFace).

A.2 Data Curation

Data Sources and Permissions: The PrimateFace dataset contains over 260,000 primate images spanning humans and non-human primates. Approximately 200,000 images were aggregated and

re-annotated from publicly available scientific datasets and repositories, utilized in accordance with their respective terms of use and licensing. An additional 60,000 images were obtained through targeted web searches of publicly accessible primate imagery. Web collection excluded copyrighted material not explicitly licensed for reuse.

Newly Collected Non-Human Primate Data: The novel collection of 60,000 non-human primate images, designed to enhance taxonomic balance within PrimateFace, was sourced from existing internal laboratory archives (where ethical approvals for original data collection were already in place), from public domain sources, and from collaborator contributions (for which they held relevant IACUC approvals). Specifically, the high-resolution video data of rhesus macaques (Macaca mulatta) used for the behavioral syllable discovery demonstration in (Figure 2) were collected at the University of Pennsylvania. These procedures were approved by the Institutional Animal Care and Use Committee (IACUC) of the University of Pennsylvania and were performed in accordance with all relevant institutional and national guidelines and regulations for the ethical treatment of animals.

Use of Human Imagery for Model Benchmarking and Demonstrations: While the openly released PrimateFace dataset focuses on non-human primates, human face data was utilized in specific contexts to demonstrate cross-species model generalization. Our models were benchmarked against standard human face datasets (WIDERFace(Yang and others [2016]), COCO-WholeBody-Face (Jin and others [2020])) – established computer vision benchmarks sourced from public repositories where ethical considerations and consent were managed by the original data creators. Additionally, the gaze-following demonstration utilized video of human infants obtained from Adobe stock footage, with no direct human subject involvement. No human images are included in the distributed PrimateFace dataset.

Data Release. The PrimateFace dataset comprises re-annotated images from existing publicly available research datasets and public domain sources. The dataset is intended to advance understanding of primate behavior and support non-invasive research methodologies that reduce the need for additional animal studies.

PrimateFace Dataset Construction. All images underwent re-annotation to conform to PrimateFace's standardized labeling protocols. Additional images were acquired through systematic web searches targeting publicly accessible content from academic institutions, research databases, and zoological society websites. To address taxonomic underrepresentation across the primate order, approximately 60,000 additional images were curated to achieve balanced representation across 60+ primate genera. This curation process set targets of at least 700 images per genus, with particular emphasis on prosimian and New World monkey taxa that are typically underrepresented in computer vision datasets. Image selection prioritized clear facial views, diverse individuals, and varied environmental contexts to maximize training data utility. The resulting PrimateFace dataset encompasses 260,000+ images representing 60+ primate genera. The collection captures diverse real-world conditions including natural habitats, zoological settings, and research contexts from previously published studies, with varied lighting, age ranges from infants to adults, multiple expressions, and social interactions (Figure 1). This comprehensive scope provides a robust foundation for cross-species facial analysis.

A.3 PrimateFace pretrained models enable high-performance, cross-species face detection

Accurate face localization across primates represents a fundamental challenge that existing approaches have yet to solve comprehensively. While existing methods achieve species-specific detection, none have demonstrated reliable performance spanning the full taxonomic breadth. PrimateFace-trained detectors address this gap directly, achieving strong face detection across all six primate superfamilies: Lemuroidea, Lorisoidea, Tarsioidea, Ceboidea, Cercopithecoidea, and Hominoidea. We trained multiple state-of-the-art object detection architectures on PrimateFace face bounding box data using default hyperparameters. Models are made available for immediate use through frameworks including MMDetection and Ultralytics.

A.3.1 PrimateFace detectors outperform human-face and open-vocabulary models on primate faces

To evaluate the face detectors, we first qualitatively compared PrimateFace-trained detectors against two alternative strategies: 1) human face detectors applied directly to primate images, and 2) open-

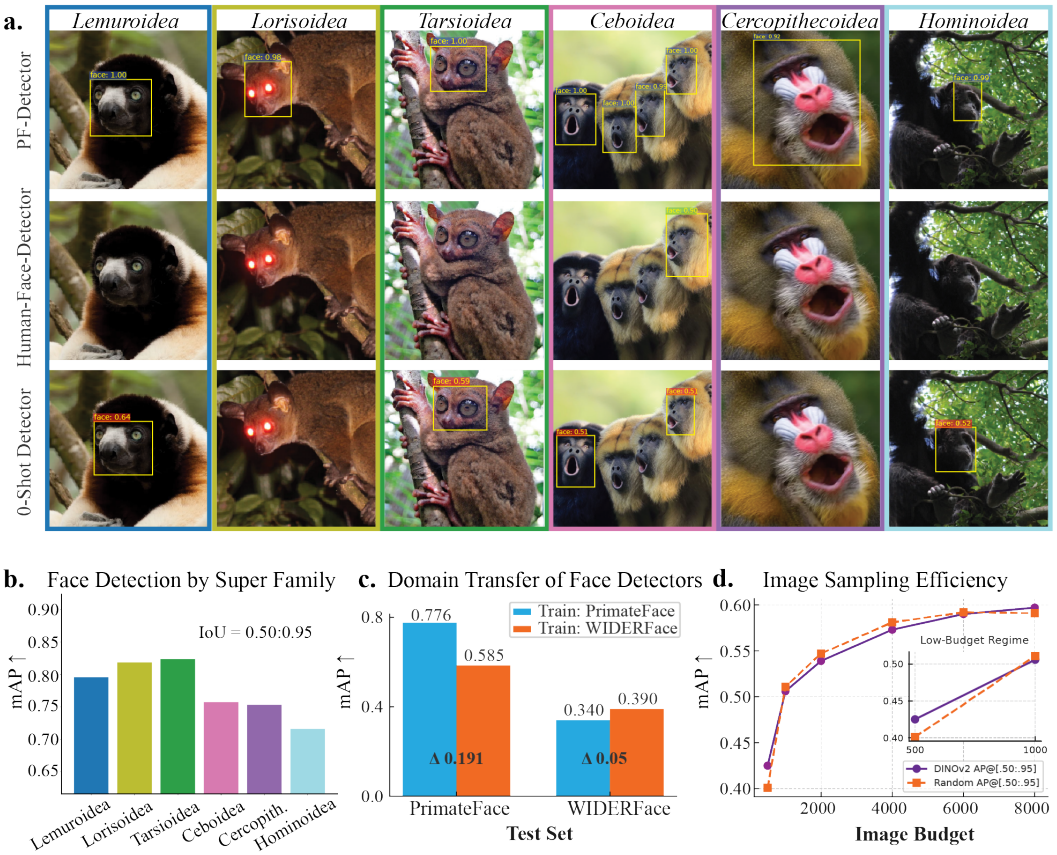

Figure A.2: **Evaluation of Face Detection Models.** (A) Qualitative comparison showing the superior performance of our PrimateFace-trained detector (top row) over a standard human-face detector (middle) and a zero-shot detector (bottom). (B) Detection performance (mAP) is robust across all six primate superfamilies. (C) Our PrimateFace-trained model shows strong zero-shot generalization to the human WIDERFace benchmark, while the human-trained model fails to generalize to our primate test set.

vocabulary detection models that can be prompted with arbitrary object categories. Open-vocabulary (or zero-shot) models like GroundingDINO (Liu et al. [2025]) represent an emerging paradigm in computer vision, using vision-language pretraining to detect objects specified through natural language prompts (e.g., 'spider monkey face') without requiring specialized training. Despite their impressive performance across many domains, these zero-shot models have yet to solve primate face detection, likely due to insufficient representation of in-the-wild primate imagery.

Qualitative comparison across these approaches reveals the advantages of our broad cross-species training approach (Figure A.2). While PrimateFace-trained models consistently detect faces across diverse primate species, human face detectors fail on many non-human primate images, and the open-vocabulary GroundingDINO detector shows inconsistent performance when prompted with "face." Despite the computer vision field's increasing emphasis on foundation models capable of zero-shot inference, comprehensive training on morphologically diverse primate facial data remains essential for reliable interspecific detection.

Beyond qualitative assessment, quantitative evaluation across primate superfamilies confirmed PrimateFace's species-agnostic performance. Models achieved highest accuracy on Tarsioidea and lowest on Hominoidea (Figure A.2), though all genera maintained detection performance (mAP @0.50:0.95) above 0.60, indicating reliable detection across taxa. Notably, even this lowest performance remained stable, as demonstrated by our models' competitive performance on human face benchmarks. This variation likely reflects differences in data curation and image contexts rather than inherent face detection difficulty: images of Tarsier monkeys typically feature subjects prominently

centered with clear facial visibility, whereas Cercopithecoidea and Hominoidea data encompass far greater diversity – from controlled laboratory settings to challenging naturalistic environments with variable lighting, poses, and occlusions.

Cross-domain evaluation on WIDERFace (Yang and others [2016]) revealed that training on diverse primate faces creates generalizable detectors. PrimateFace-trained models achieved 0.340 mAP@0.50 on the challenging WIDERFace human benchmark – within 0.05 of models trained specifically on human faces (0.390 mAP@0.50). In contrast, human-face detectors showed dramatic performance degradation when applied to primate faces, achieving only 0.585 mAP on our PrimateFace test set compared to 0.775 mAP for PrimateFace-trained detectors (Figure A.2). This disparity demonstrates that PrimateFace's inherent morphological heterogeneity provides a richer training foundation that transfers effectively across species, while human-centric training captures only a narrow slice of possible facial variation, suggesting cross-primate variation may provide natural data augmentation that enhances generalization.

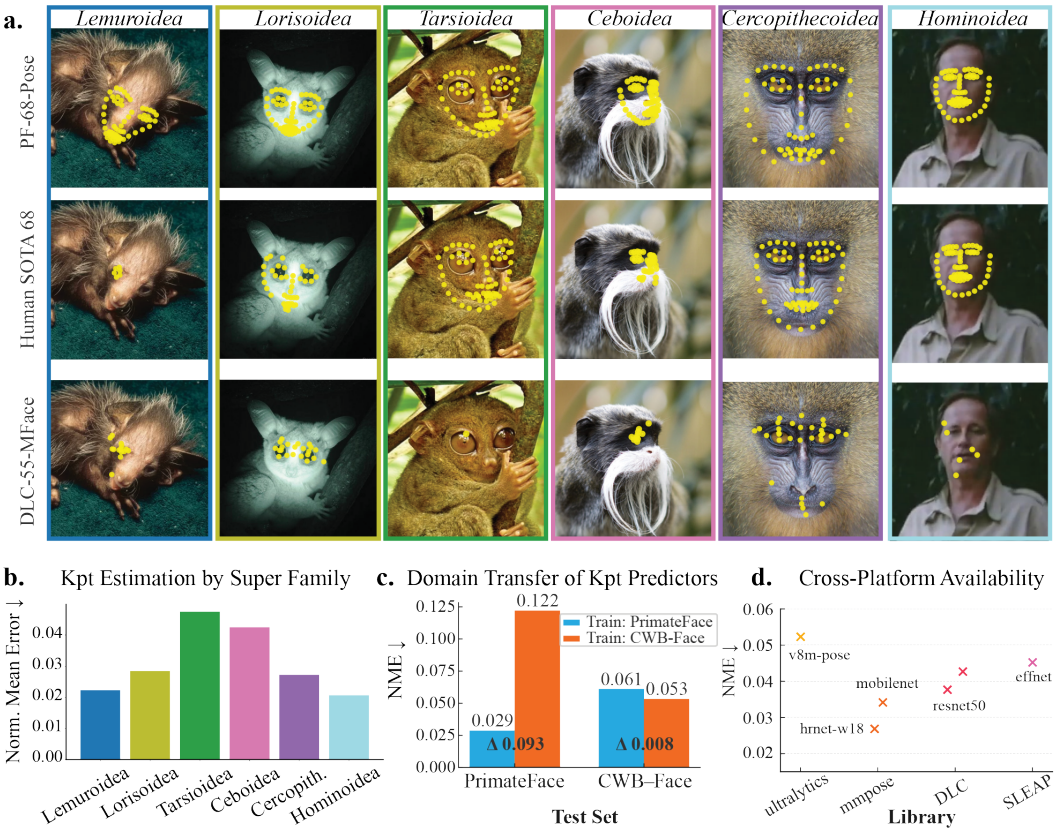

Figure A.3: **Evaluation of Facial Landmark Estimation (FLE) Models.** (A) Qualitative comparison showing our PrimateFace-trained model (top row) accurately localizes landmarks where human-specific (middle) and macaque-specific (bottom) models fail. (B) Normalized Mean Error (NME) is consistently low across superfamilies. (C) Our model generalizes to the human COCO-WholeBody-Face benchmark with performance competitive to a specialist model. This generalization is asymmetric, as the human-trained model performs poorly on our primate test set, highlighting the benefit of diverse pretraining.

A.4 PrimateFace facilitates species-agnostic, accurate facial landmark estimation

Facial landmark localization enables detailed analysis of primate expressions, gaze patterns, and behavioral kinematics by providing consistent reference points that can be tracked over time in video sequences. When applied frame-by-frame to videos, these landmarks quantify dynamic facial movements – such as mouth opening during vocalizations, eye gaze direction, or brow movements during social interactions – providing objective kinematic measurements for behavioral analysis

that previously relied on subjective manual assessment. While existing NHP landmark models like DeepLabCut's MacaqueFace represent pioneering efforts for individual species, no solution has achieved reliable landmark estimation spanning the full primate order. PrimateFace-trained models address this limitation, delivering robust 68-keypoint facial landmark estimation across all six primate superfamilies using top-down pose estimation architectures that build upon face detection outputs (Figure A.3). We trained multiple state-of-the-art pose estimation models on the PrimateFace dataset using default hyperparameters, interoperable with platforms including MMPose, Ultralytics, DeepLabCut, and SLEAP for seamless integration.

A.4.1 PrimateFace landmarks outperform human and species-specific models across primate faces

To evaluate landmark estimation performance, we compared PrimateFace-trained models against human facial landmark models and existing primate-specific approaches. Our models employ a top-down pose estimation framework that first detects faces and then estimates landmarks within bounding boxes, enabling flexible integration with other frameworks (e.g., Ultralytics or mmdetection). For comparison, we evaluated human-face landmark models and DeepLabCut's species-specific MacaqueFace model. While MacaqueFace uses a bottom-up approach that directly detects keypoints, we provided cropped face bounding boxes to optimize its performance, effectively creating similar input conditions. Across frameworks, top-down models trained on PrimateFace data perform comparably when using consistent training protocols.

Qualitative comparison of landmark estimation approaches reveals the advantages of cross-species training (Figure A.3). PrimateFace models consistently localize facial landmarks across diverse primate genera, while human-face models show systematic failure modes – particularly misinterpreting mouth contours as facial boundaries in Cercopithecoidea, leading to cascading errors in nose and mouth landmark placement (see Fig. 5a, 2nd row, 5th column). Human models completely fail on morphologically distant species like Lemuroidea, as seen with Daubentonia. The DeepLabCut MacaqueFace-55 model occasionally performs well but demonstrates the limitations of species-specific training when applied beyond its target domain.

A.4.2 Facial landmark estimation maintains consistent performance across primate superfamilies

Quantitative evaluation across primate superfamilies revealed performance patterns reflecting both data availability and morphological constraints. Models exhibited lowest performance on Tarsioidea due to limited training data and their distinctive facial morphology – characterized by proportionally enormous eyes and compressed facial features that differ substantially from other primates. PrimateFace landmark estimators achieved optimal results on Hominoidea and Lemuroidea (Figure A.3). Performance variation across genera reflects multiple factors: high-performing genera (Macaca, Pan) benefit from extensive datasets, both within and beyond the laboratory setting, and share more typical primate facial proportions, while challenging cases (Alouatta, the Howler Monkey) involve rapid facial movements during vocalizations that stress landmark tracking algorithms. This breakdown demonstrates the value of taxonomically-informed evaluation and highlights both data collection priorities and morphological considerations for future dataset expansions.

A.4.3 PrimateFace models maintain human facial landmark estimation performance

Cross-domain evaluation on COCO-WholeBody-Face demonstrated the generalization advantages of morphologically diverse training. PrimateFace-trained models achieved 0.061 normalized mean error (NME) on human landmark estimation – within 0.008 of human-face baselines (0.053 NME) (Figure A.3). In contrast, human-face models showed substantial degradation on primate faces (0.122 vs. 0.029 NME on PrimateFace data, delta 0.093). This reinforces that cross-species training provides feature representations that transfer effectively to humans, while human-centric training fails to capture broader facial variation.

This holds implications for comparative primatology and translational research. Researchers can now apply consistent landmark schemes across species for evolutionary studies, employ unified analysis pipelines for mixed human-NHP datasets, and leverage the growing ecosystem of human facial analysis tools for non-human primate behavioral research regardless of existing computational infrastructure.