ANIMAL WORLD: A CROSS-SPECIES DATASET FOR SOCIAL ANIMAL POSE UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Animal pose understanding in the wild remains challenging due to large morphological variation, frequent occlusion, and complex social interactions. Reliable 2D pose estimation is not only fundamental to animal image understanding, but also provides an important source of supervision and evaluation for monocular 3D animal reconstruction, where accurate real-world 3D ground truth remains difficult to obtain. Existing animal pose benchmarks, however, are typically limited to particular species, sparse pose definitions, or isolated individuals, leaving robustness of animal pose estimation under diverse natural scenes insufficiently characterised. We introduce Animal World, a scene-centric dataset and benchmark for cross-species 2D animal pose understanding that jointly captures morphological diversity and inter-animal interaction in natural environments. Animal World provides a unified anatomical keypoint taxonomy and large-scale instance-level pose annotations, which enable consistent evaluation across species and socially complex scenes. Building on Animal World, we establish a comprehensive diagnostic benchmark that systematically evaluates the generalisation and robustness of animal pose estimation under species and domain shifts, occlusion, and diverse social configurations. Our extensive evaluations reveal substantial performance variation across species, domains, and interaction contexts, highlighting challenges beyond conventional single-animal benchmarks. We further show that Animal World supervision provides more informative 2D pose cues for downstream 3D animal reconstruction, demonstrating its value beyond 2D pose estimation alone. Animal World provides a unified data and evaluation foundation for advancing robust cross-species animal understanding in complex natural environments.

1 INTRODUCTION

Animals are central to natural ecosystems, and their behavior provides important cues for understanding social organization, ecological adaptation, species interaction, and environmental change. However, animal pose understanding remains substantially less developed than its human counterpart. Unlike humans, for whom large-scale datasets and thousands of scans can be collected under controlled conditions (Ionescu et al., 2013; Huang et al., 2022; Yu et al., 2020), animals are non-cooperative, highly diverse in morphology, and difficult to capture systematically in laboratory settings. Their body structure and articulation also vary substantially across species, from big cats and ungulates to bears, elephants, and primates.

Existing animal pose datasets have made substantial progress toward broader species coverage. Early benchmarks often focus on individual species (Biggs et al., 2020; Mathis et al., 2021; Badger et al., 2020; Joska et al., 2021), while more recent datasets cover a wider range of animal categories and pose variations (Xu et al., 2023; Ng et al., 2022; Yu et al., 2021). However, most benchmarks remain organized around individual animal instances or relatively clean pose observations. In natural in-the-wild environments, animals frequently appear in groups, physically interact, overlap with one another, or co-occur with other species. These settings introduce challenges beyond isolated keypoint localization, including strong mutual occlusion, ambiguous instance association, heterogeneous morphology within the same scene, and interaction-driven articulation. Although existing datasets may contain multi-animal images, they are not designed to systematically evaluate pose estimation in scene-centric multi-animal settings (Yu et al., 2021; Yang et al., 2022b). As a result, the robustness of animal pose estimation models under diverse natural social contexts remains in-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

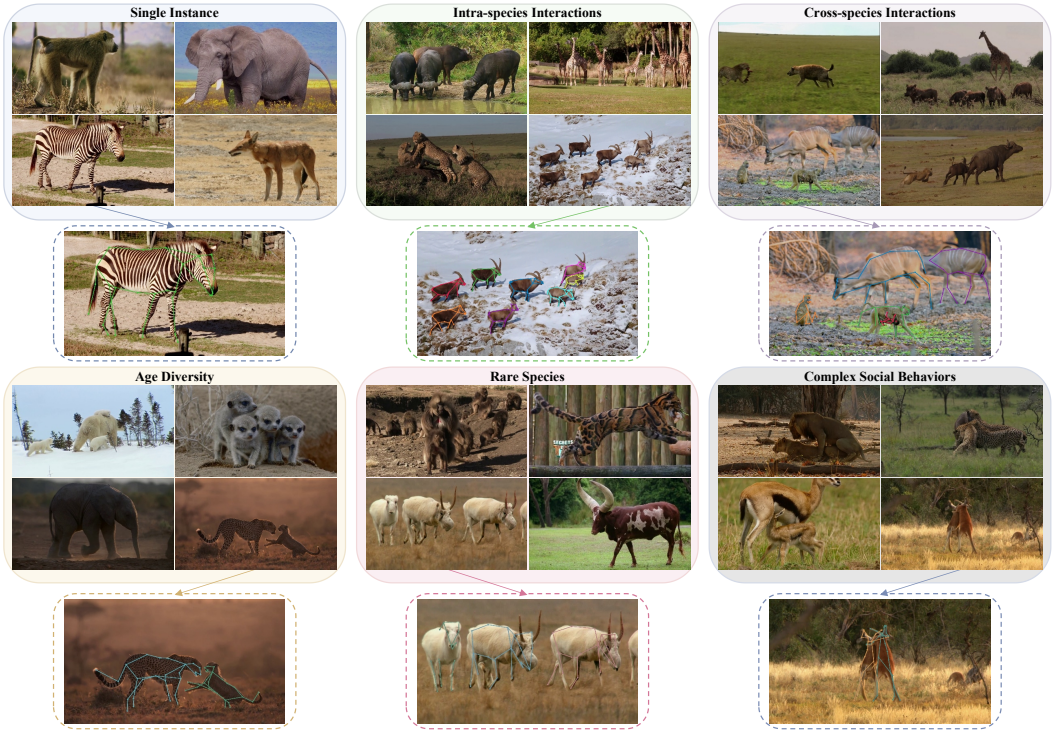


Figure 1: Overview of Animal World. Animal World covers diverse animal pose understanding scenarios in the wild, including single instances, intra-species interactions, cross-species interactions, age diversity, rare species, and complex social behaviors. For each category, we show representative source images together with zoomed examples containing instance-level pose annotations.

sufficiently characterized. Meanwhile, 2D keypoints remain an important source of supervision and evaluation for animal mesh recovery on real imagery. Advancing real-world 2D pose benchmarks can therefore also provide more reliable anatomical cues for the development and evaluation of 3D animal reconstruction methods (Yu et al., 2026).

To make progress, we introduce **Animal World**, a large-scale, scene-centric dataset for cross-species 2D animal pose understanding in the wild. By scene-centric, we mean that the dataset preserves and explicitly characterizes the composition of the full animal scene, rather than treating each animal as an independent pose instance. This includes how many animals are present, which species co-occur, whether they interact, and how these factors induce occlusion, contact, and unusual articulation. As a result, Animal World jointly captures broad morphological diversity and inter-animal interaction in natural scenes, extending conventional animal pose benchmarks beyond isolated, single-animal settings. We define a unified 30-keypoint anatomical taxonomy and provide instance-level masks and bounding boxes to support fine-grained and consistent pose evaluation across diverse species and socially complex scenes. This unified representation enables models to be studied under a common anatomical space despite substantial differences in animal morphology.

Beyond providing a new dataset, Animal World is designed as a diagnostic benchmark for studying the generalization and robustness of animal pose estimation models. We systematically evaluate representative architectures under in-domain evaluation, cross-family generalization, cross-dataset transfer, occlusion, social configurations, and other challenging natural-scene conditions. Our experiments reveal substantial performance variation across species, domains, and interaction contexts, highlighting challenges that are not exposed by conventional single-animal evaluation. We further investigate the utility of Animal World for 3D animal reconstruction and show that models trained with its dense 2D anatomical supervision can provide more informative keypoint cues for downstream mesh recovery. Together, Animal World is established as a unified data and evaluation foundation for robust cross-species animal pose understanding in complex natural environments.

Our main contributions are threefold. First, we introduce **Animal World**, a large-scale, scene-centric dataset and benchmark for cross-species 2D animal pose understanding in the wild, extending evaluation beyond conventional isolated, single-animal settings through broad morphological diversity and inter-animal interactions. Second, we define a unified 30-keypoint anatomical taxonomy, together with instance-level masks and bounding boxes, to support fine-grained and consistent 2D pose evaluation across diverse species and complex natural scenes. Finally, we establish a comprehensive diagnostic benchmark for evaluating generalization and robustness across species, domains, and challenging scene conditions, and further show that Animal World supervision provides informative 2D anatomical cues for downstream 3D animal reconstruction.

2 Related Work

2.1 Species-specific Animal Pose Datasets

Compared to large-scale human pose datasets such as MS COCO (Lin et al., 2014) and CrowdPose (Li et al., 2019a), animal pose datasets remain comparatively limited in scale, taxonomic coverage, and annotation consistency. Collecting animal data at scale is inherently challenging for two reasons. First, animals are non-cooperative and hard to be systematically recorded under controlled conditions. Second, they exhibit substantial inter-species variation, making large-scale data collection and standardized annotation considerably more difficult. Consequently, early animal pose estimation datasets were mainly species-specific, e.g., horses (Mathis et al., 2021), pigs (An et al., 2023), dogs (Biggs et al., 2020; Shooter et al., 2024a), goats (Jin et al., 2026), mice (Segalin et al., 2021; Hu et al., 2023), and macaques (Bala et al., 2020; Labuguen et al., 2021), with related datasets for tigers (Li et al., 2019b) and birds (Wah et al., 2011; Van Horn et al., 2015) providing additional pose or part-level annotations. Beyond 2D poses, a complementary line of datasets studies animal motion and 4D reconstruction across different species. PFERD (Li et al., 2024a) provides multi-view 3D motion capture for horses, while DogMo (Wang et al., 2025), InterPet4D (Peng et al., 2026), and CoP3D (Sinha et al., 2023) cover canine or feline motion, interaction, and dynamic reconstruction. BigMaQ (Martini et al.) extends multi-view dynamic capture to interacting rhesus macaques. Beyond species-specific datasets, AiM (Zhao et al., 2025) provides a large-scale in-the-wild benchmark for 4D reconstruction across diverse quadrupeds. While these datasets have enabled substantial progress within individual species, their heterogeneous annotation schemes and restricted taxonomic coverage lead to fragmented evaluation across animal categories.

2.2 Cross-species Animal Pose Datasets

To support broader cross-species animal understanding, several datasets extend pose and shape annotations across multiple species. Animal Pose (Cao et al., 2019) and AwA-Pose (Banik et al., 2021) provide 2D keypoint annotations for 5 and 35 species, respectively, while AP-10K (Yu et al., 2021) introduces a unified 17-keypoint taxonomy over 54 species. APT-36K (Yang et al., 2022b) extends large-scale 2D pose estimation to video and tracking across 30 species, and Animal Kingdom (Ng et al., 2022) further broadens taxonomic coverage to hundreds of species for pose and behavior understanding. There are also 3D dataset cover multiple species. Animal3D (Xu et al., 2023) provides multi-species 3D pose and shape annotations, while WildDepth (Aamir et al., 2026) introduces synchronized RGB–LiDAR observations for wildlife 3D perception. Despite their broad species coverage, these datasets remain primarily organized around individual animal instances or conventional pose and shape estimation, providing limited support for systematic scene-centric multi-animal evaluation. Synthetic datasets (Choi et al., 2026; Lyu et al., 2025; Shooter et al., 2024b; Niewiadomski et al., 2025; Hu et al., 2026) provide scalable and precise 3D supervision that is difficult to obtain in the wild, but generalization from synthetic data to the morphological and scene complexity of real-world animals remains challenging.

2.3 3D Animal Reconstruction Methods

A typical application of 2D poses is guiding 3D animal reconstruction, especially parametric template based methods. SMAL (Zuffi et al., 2017) introduces a skinned parametric model for representing the 3D shape and pose of diverse quadrupeds. Subsequent SMAL-based methods, including SMALR (Zuffi et al., 2018), WLDO (Biggs et al., 2020), BARC (Rüegg et al., 2022), BITE (Rüegg



Figure 2: Keypoint visualization.

ID	Name	ID	Name
0	Left eye	15	Right forelimb wrist
1	Right eye	16	Left hind-limb ankle
2	Lower jaw	17	Right hind-limb ankle
3	Left forefoot	18	Neck midpoint
4	Right forefoot	19	Tail tip
5	Left hind foot	20	Left ear base
6	Right hind foot	21	Right ear base
7	Tail root	22	Left mouth corner
8	Left forelimb elbow	23	Right mouth corner
9	Right forelimb elbow	24	Nose tip
10	Left hind-limb knee	25	Tail midpoint
11	Right hind-limb knee	26	Anterior back
12	Left upper forelimb	27	Middle back
13	Right upper forelimb	28	Posterior back
14	Left forelimb wrist	29	Abdomen midpoint

Table 1: Definition of the 30 animal keypoints.

et al., 2023), and SMAL+ (Zuffi & Black, 2024), extend parametric template-based reconstruction through richer species or breed-specific shape priors and weak image supervision. Coarse-to-fine reconstruction (Li & Lee, 2021) further combines SMAL estimation with non-parametric mesh deformation to improve shape expressiveness. In parallel, LASR (Yang et al., 2021), BANMo (Yang et al., 2022a), LASSIE (Yao et al., 2022), MagicPony (Wu et al., 2023), SAOR (Aygun & Mac Aodha, 2024), 3D-Fauna (Li et al., 2024b), Ponymation (Sun et al., 2024), and BAT3R (Zadrozny et al., 2026) explore weakly supervised or template-free learning of articulated and animatable animal representations from image collections or monocular videos.

Recent work further emphasizes scalable multi-species supervision and generalization. AniMer (Lyu et al., 2025) and AniMer+ (An et al., 2026) combine real and synthetic data for unified multi-species reconstruction. Generative Zoo (Niewiadomski et al., 2025) scales synthetic 3D supervision and demonstrates strong transfer to real images. PRIMA (Yu et al., 2026) incorporates biological priors, pseudo-3D supervision, and test-time adaptation for cross-species mesh recovery, whereas SAM3D Animal (Hu et al., 2026) extends single-instance reconstruction to promptable multi-animal settings. In our paper, we will show how denser keypoints improve 3D reconstruction with RPIMA.

3 Dataset

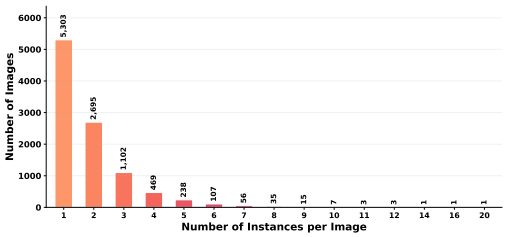
3.1 Dataset Description

Data Collection. We construct Animal World from more than 500 in-the-wild wildlife long videos collected from YouTube. Although the raw materials are videotapes, we do not organize the dataset as continuous video sequences because we focus on pose estimation instead of motion analysis. Therefore, we sparsely sample frames with a large temporal stride to reduce near-duplicate content and increase visual diversity across poses, species, viewpoints, and social configurations. Note that, we treat all samples as independent images throughout the paper. This design allows Animal World to capture diverse social animal scenes while avoiding the redundancy commonly introduced by densely sampled video frames.

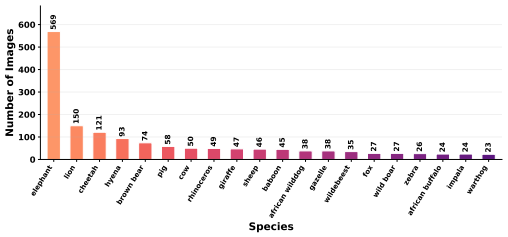
Scene-Centric Diversity. Animal World is designed to capture the morphological and interaction complexity of animals in different natural scenes. The dataset annotates both single-species groups and cross-species co-occurrence. For images with multiple species, the species labels are assigned at the instance level while image-level composition records mixed-species scenes. Such settings introduce substantial variation in body size, shape, texture, and articulation within the same image, often together with mutual occlusion and ambiguous instance association. The dataset also covers animals across different life stages, including adults, juveniles, and parent-offspring groups, introducing age-dependent morphological variation that is often underrepresented in existing animal benchmarks. These examples provide a basis for studying whether pose representations remain reliable across changes in body proportion and appearance, while also offering useful cues for downstream 3D reconstruction, where juvenile morphology can be difficult to recover (Zuffi & Black, 2024). Finally, Animal World includes diverse social interactions, such as chasing, fighting, grooming, mating, parental care, group movement, and collective foraging. These scenes couple pose estimation with body contact, unusual articulation, mutual occlusion, and close inter-animal interaction, enabling evaluation beyond conventional isolated-animal settings.



(a) Taxonomic distribution. The inner circle presents different families, while the outer circle presents more detailed species belonging to the animal family. We recommend scrolling the mouse for better zoom-in visualization.



(b) A detailed breakdown of instance-count distribution per image in Animal World.



(c) A detailed breakdown of the number of images collected in Animal World for juveniles of each animal species

Figure 3: **Overview of Animal World.** (a) Taxonomic distribution of species and families in Animal World. (b) Distribution of the number of annotated animal instances per image. (c) Distribution of age-diverse images across juvenile animal species.

3.2 Dataset Statistics

Total Number and Annotation Type. Animal World contains 10,036 labelled images and 19,132 annotated animal instances, covering 124 species with a unified 30-keypoint taxonomy manually annotated by experts. The 30 anatomical keypoints are unified for both quadrupeds and primates, as shown in Fig. 2 and Table 1. They cover five anatomical regions—the head, limbs, tail, dorsal body, and ventral body—including major facial landmarks, limb joints, tail points, and trunk landmarks. This denser and anatomically structured representation supports fine-grained and consistent pose evaluation across diverse animal species. Each instance is further associated with a SAM3 (Carion et al., 2025)-derived segmentation mask and bounding box, which are manually verified to ensure annotation quality and correctness. Qwen3VL-8B (Bai et al., 2025) provides instance-level behavior descriptions for all 19,132 instances and image-level group social interaction labels for all 10,036 images; 60.9% of multi-animal images have a group-interaction label.

Taxonomy. As shown in Fig. 3(a), Animal World spans a broad taxonomic distribution across diverse animal families, providing cross-species coverage for animal pose understanding in the wild. A key property of Animal World is its explicit emphasis on multi-animal scenes. Fig. 3(b) shows the distribution of the number of annotated instances per image. While 5,016 images contain a single animal instance, the remaining 5,020 images contain multiple animals, ranging from 2 to 20 annotated instances per image and accounting for 50.02% of the dataset. This distribution enables evaluation under crowded animal groups, mutual occlusion, and complex instance association. Animal World also includes 489 cross-species images, covering 97 unique species combinations. These images are designed to test whether pose estimation models can remain robust when animals with different body sizes, shapes, textures, and articulation patterns appear in the same visual context.

Age-Diverse Images. The dataset contains 1,918 age-diverse images, in which only juvenile animals exist, or both juvenile and adult animals appear together. Fig. 3(c) summarizes the distribution of juvenile examples across species. This subset supports evaluation of age-sensitive animal pose understanding, where models must distinguish juvenile-specific morphology from simple scale or depth changes.

Table 2: Comparison with previous datasets. "Mixed-species Comb.": the number of distinct species combinations. "Rare Inst.": the number of rare-species animals. "Age-diverse Inst.": the numbers of juvenile instances. A dash indicates that the annotation or statistic is not available.

Dataset	Labelled Images/Instances	Instances	Species	Keypoints	Intra-species Groups	Mixed-speices Comb.	Social Interactions	Rare Inst.	Age-diverse Inst.
Animal Pose Cao et al. (2019)	4,666	6,117	5	20	Yes	–	No	–	–
StanfordExtra Biggs et al. (2020)	12,000	12,000	1	20	No	–	No	–	–
Horse-30 Mathis et al. (2021)	8,114	8,114	1	22	No	–	No	–	–
AwA-Pose Banik et al. (2021)	10,064	10,064	35	39	No	–	No	–	–
AP-10K Yu et al. (2021)	10,015	13,028	54	17	Yes	–	No	1,244	–
Animal Kingdom Ng et al. (2022)	33,099	33,099	850	23	No	–	No	428	–
APT-36K Yang et al. (2022b)	36,000	53,006	30	17	Yes	–	No	9,728	–
Animal3D Xu et al. (2023)	3,400	3,400	40	26	No	–	No	–	–
AiM Zhao et al. (2025)	11,000	11,000	23	17	No	–	No	–	–
Animal World (ours)	**10,036**	**19,132**	**124**	**30**	**Yes**	**97**	**Yes**	**4,407**	**4,499**

Rare-species definition. We label an image as rare if any of its constituent taxa is listed by the IUCN Red List as Critically Endangered (CR) or Endangered (EN). Under this rule, rare images comprise $1{,}896$ (21.0%) of the training set and 211 (21.0%) of the validation set.

Difference from Existing Datasets. Table 2 compares Animal World with representative public animal pose estimation datasets. While existing benchmarks substantially broaden species coverage, their pose evaluation is largely centered on individual instances or temporal tracking. Animal Kingdom (Ng et al., 2022) provides broad taxonomic diversity but remains primarily instance-centric, whereas APT-36K (Yang et al., 2022b) emphasizes pose estimation and temporal association in video. Animal World instead places scene-centric multi-animal imagery at the core of both data organization and evaluation. Together with a unified 30-keypoint taxonomy, this design enables systematic analysis of cross-species generalization and pose estimation robustness under diverse morphology, inter-animal interaction, and complex natural scenes.

4 Experiments

Dataset Split. Animal World contains 9,032 training images with 17,258 annotated instances and 1,004 validation images with 1,874 instances, an approximately 90:10 image-level split. Given the long-tailed distribution of animal categories, the dataset split retains most annotated examples for training while reserving held-out images for validation.

Baselines. To establish a comprehensive benchmark for animal pose estimation on the proposed Animal World dataset, we evaluate four representative models (HRNet Sun et al. (2019), HRFormer Yuan et al. (2021), ViTPose Xu et al. (2022), and RTMPose Jiang et al. (2023)) conditioned on ground-truth bounding boxes. For the 3D animal reconstruction task, following Lyu et al. (2026), we perform post-optimization on PRIMA Yu et al. (2026). Specifically, we employ 2D keypoints predicted by a ViTPose model—trained on the Animal World dataset—as supervisory signals. We then compare this optimized framework against the baseline PRIMA Yu et al. (2026).

Evaluation Metrics. For the animal pose estimation, we adopt Average Precision (AP) Lin et al. (2014) as main metric. Specifically, we use AP_{50} where OKS threshold is set to 0.5, and AP which is the average of precision where OKS ranges from 0.5 to 0.95 with 0.05 interval. Furthermore, to assess 3D animal reconstruction accuracy, we utilize the Percentage of Correct Keypoints (PCK), AP_{50}, and AP.

Implementation Details. All animal pose estimation models are implemented within the MMPose Contributors (2020) framework. Consistent with standard practices in this domain, the models are trained for 210 epochs. We set the initial learning rate to 5e-4, applying a tenfold decay at epochs 170 and 200.

4.1 Animal Pose Estimation

Table 3 establishes the multi-backbone baseline on Animal World. All four architectures achieve strong in-domain performance, with ViTPose-B obtaining the highest mAP across all evaluation settings. When evaluated on APTv2 using the shared 16 keypoints, performance consistently decreases

across backbones, indicating a clear cross-dataset domain shift. On Animal World, extending evaluation from the shared 16 keypoints to the full 30-keypoint taxonomy reduces mAP by 1.8–2.9 points while leaving AP unchanged, which suggests that the denser annotation primarily increases fine-grained localization difficulty. Overall, the consistent trends across architectures indicate that the observed challenges are not specific to a particular backbone.

Table 3: **Multi-backbone benchmark trained on Animal World.** We train representative pose-estimation backbones on Animal World and evaluate them on APTv2 and Animal World. APTv2 and Animal World (shared) use the 16 semantically matched keypoints. Animal World (30 kpt) evaluates the native Animal World keypoint set.

Dataset	HRNet-w48		HRFormer-B		ViTPose-B		RTMPose	
	AP_{50}	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}	AP
APTv2 (shared 16)	88.3	58.8	86.5	56.5	88.1	59.6	88.0	56.5
Animal World (shared 16)	96.9	78.6	95.9	78.7	96.9	80.1	95.9	76.3
Animal World (full)	96.9	76.3	95.9	76.8	96.9	78.3	95.9	73.4

4.2 Occlusion Evaluation

We evaluate robustness to occlusion by grouping instances according to the number of visible keypoints, using visibility as a proxy for occlusion severity. As shown in Fig. 4, ViTPose-B models trained with Animal World supervision consistently outperform the APTv2-trained baseline on Animal World under low-visibility regimes, with the full 30-keypoint supervision showing the strongest advantage under severe occlusion. As more keypoints become visible, the performance gap progressively narrows, indicating that the benefit of Animal World is concentrated in challenging partially observed cases. On APTv2 (Fig. 5), the in-domain model retains an advantage under severe occlusion, while Animal World-trained models become competitive as visibility increases. These results suggest that the scene-centric supervision in Animal World improves robustness to the occlusion patterns encountered in complex natural animal scenes.

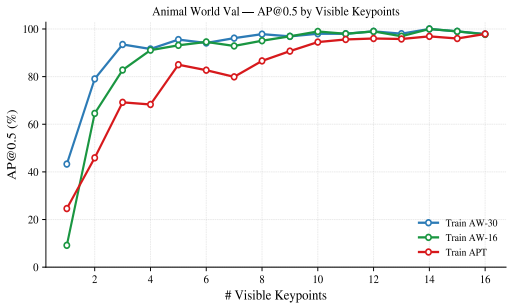


Figure 4: Performance of ViTPose-B models trained with AW-30, AW-16, and APTv2 supervision and evaluated on Animal World.

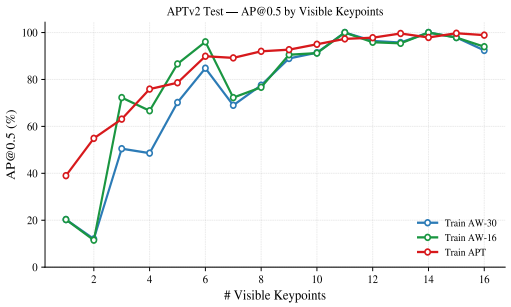


Figure 5: Performance of ViTPose-B models trained with AW-30, AW-16, and APTv2 supervision and evaluated on APTv2.

4.3 Impact of Social Interaction Scene Supervision

Table 4 examines the contribution of social interaction training scenes. Adding social-scene supervision has only a modest effect on single-animal images (+1.3 mAP), but yields substantially larger gains on intra-species groups (+7.7), cross-species scenes (+5.3), and social interactions (+9.2), improving overall mAP by 5.7 points. Social-only training already recovers much of the performance on interaction-heavy subsets, while the full training set consistently performs best, indicating that interaction-rich and conventional examples provide complementary supervision. These results support the role of scene-centric training data in improving pose robustness under complex inter-animal interactions.

Table 4: **Effect of social-scene supervision on Animal World.** We evaluate models trained with different data regimes on scene-centric Animal World test subsets and the full evaluation set. The comparison between Animal World w/o Social and Animal World Full measures the contribution of social-scene supervision.

Training Data	Single		Intra-species Group		Cross-species		Social		Full Eval	
	AP_{50} ↑	AP ↑	AP_{50} ↑	AP ↑	AP_{50} ↑	AP ↑	AP_{50} ↑	AP ↑	AP_{50} ↑	AP ↑
Animal World w/o Social	99.0	83.8	91.3	67.5	92.1	76.4	88.9	64.8	93.5	72.6
Animal World Social-only	98.0	78.3	95.7	72.5	96.7	79.8	94.7	71.6	96.8	74.6
Animal World Full	99.0	85.1	95.8	75.2	96.7	81.7	94.7	74.0	96.9	78.3
Gain over w/o Social	+0.0	+1.3	+4.5	+7.7	+4.6	+5.3	+5.8	+9.2	+3.4	+5.7

Table 5: **Effect of age-diverse supervision on pose estimation.** We compare models trained without age-diverse examples, on age-diverse examples only, and on the full Animal World training set across age-related evaluation subsets.

Training Data	Juvenile-only		Family		Age-diverse		Full Eval	
	AP_{50} ↑	AP ↑	AP_{50} ↑	AP ↑	AP_{50} ↑	AP ↑	AP_{50} ↑	AP ↑
Animal World w/o Age	97.0	73.5	97.8	72.7	97.9	73.0	96.9	77.0
Animal World Age-only	100.0	75.0	97.9	71.9	97.9	72.6	95.8	71.5
Animal World Full	97.0	78.2	97.9	75.6	97.9	76.1	96.9	78.3
Gain over w/o Age	+0.0	+4.7	+0.1	+2.9	+0.0	+3.1	+0.0	+1.3

4.4 Impact of Age-Diverse Supervision

Table 5 evaluates the effect of age-diverse supervision on pose estimation. Compared with training without age-diverse examples, the full Animal World training set improves mAP by 4.7 points on juvenile-only animals and by 2.9–3.1 points on family and age-diverse scenes, while yielding a smaller 1.3-point gain on the full evaluation set. The larger improvements on age-related subsets indicate that exposure to diverse life stages is particularly beneficial for fine-grained localization under age-dependent morphological variation. Training only on age-diverse examples does not match the full training set, suggesting that age-specific supervision is most effective when combined with broader morphological and scene diversity.

4.5 Impact of Rare-Species Supervision

Table 6 evaluates the contribution of rare-species supervision. Removing rare taxa from training causes a substantial drop on the rare-species subset, reducing mAP from 72.5 to 55.4, while performance on common species remains nearly unchanged. Including rare-species examples therefore yields a 17.1-point mAP improvement on rare taxa, compared with only 0.4 points on common species, and improves overall mAP by 4.3 points. The concentrated gain indicates that explicit coverage of underrepresented taxa is important for maintaining pose accuracy beyond commonly observed species.

4.6 Impact of Multi-Animal Scenes

Table 7 leverages the broad multi-animal coverage of Animal World to analyze pose robustness as scene occupancy increases. Performance drops markedly from single to multi-animal images, with mAP decreasing from 85.0 to 75.9 for two-animal scenes, while remaining nearly unchanged from 2 to 5+ animals. This pattern suggests that the main difficulty arises from the transition to multi-animal settings rather than from instance count alone. More importantly, the result highlights the value of Animal World as a scene-centric benchmark that exposes challenges largely hidden by conventional single-animal evaluation.

4.7 3D Animal Reconstruction

Table 6: **Effect of rare-species supervision on animal pose estimation.** We evaluate models trained with and without IUCN CR/EN (rare) taxa on common, rare, and full Animal World evaluation splits.

Training Data	Common		Rare		Full Eval	
	AP_{50} ↑	AP ↑	AP_{50} ↑	AP ↑	AP_{50} ↑	AP ↑
Animal World w/o Rare	96.8	79.6	89.3	55.4	94.8	74.1
Animal World Full	96.9	80.0	96.8	72.5	96.9	78.4
Gain over w/o Rare	+0.1	+0.4	+7.5	+17.1	+2.1	+4.3

Table 8: **Ablation study of different keypoint prompt sources on Animal3D and Animal World.** PRIMA-base denotes the original PRIMA Yu et al. (2026) method without keypoint prompts. ViTPose-Base uses keypoint prompts predicted by a ViTPose-Base trained on APTv2 Xu et al. (2022), while ViTPose-AW uses keypoint prompts predicted by a ViTPose-Base trained on Animal World.

Model	Animal3D			Animal World		
	PA-MPJPE↓	PCK@0.1↑	AUC↑	AUC↑	PCK@0.1↑	PCK@0.15↑
PRIMA-base	75.3	0.627	0.896	0.894	0.646	0.803
w/ ViTPose-Base	75.1	0.557	0.884	0.891	0.615	0.777
w/ ViTPose-AW	**74.4**	**0.650**	**0.901**	**0.908**	**0.705**	**0.848**

The ablation results in Tab. 8 show that keypoint prompts trained with the denser Animal-World annotation scheme provide stronger geometric guidance for PRIMA, especially under cross-dataset generalization. On Animal3D, which serves as an out-of-distribution benchmark, ViTPose-AW achieves the best overall performance, reducing PA-MPJPE from 75.4 for PRIMA-base and 75.1 for ViTPose-Base to 74.4, while also improving PCK@0.1 and AUC to 0.650 and 0.901, respectively. A similar trend is observed on AnimalWorld, where ViTPose-AW substantially outperforms both PRIMA-base and ViTPose-Base, improving AUC from 0.894/0.891 to 0.908 and PCK@0.1 from 0.646/0.615 to 0.705. These gains indicate that prompts from the Animal World-trained model provide more reliable and anatomically informative constraints, likely because the 30-keypoint supervision captures richer body structure than the 17-keypoint APTv2.

#Animals	Images	Inst.	AP_{50} ↑	AP ↑
1	501	501	99.0	85.0
2	312	624	96.9	75.9
3	100	300	96.5	75.9
4	57	228	95.5	75.4
5+	34	218	93.1	75.9

Table 7: **Pose performance by animal count.** We partition the Animal World validation set by the number of instances per image.

5 Conclusion and Limitations

We introduce **Animal World**, a large-scale, scene-centric dataset and benchmark for cross-species 2D animal pose understanding in the wild. With a unified 30-keypoint taxonomy, instance-level masks, and bounding boxes, Animal World provides a consistent annotation space for evaluating animal pose across diverse morphologies and complex multi-animal scenes. Our experiments reveal clear generalization gaps across animal families and datasets, and show that performance degrades substantially when moving from isolated to multi-animal settings. We further find that interaction-rich, age-diverse, and rare-species supervision provides targeted improvements on the corresponding challenging subsets, highlighting the value of broad scene and morphological coverage. Beyond 2D pose estimation, Animal World-trained keypoints also provide useful supervision for downstream 3D animal reconstruction. Together, these results establish Animal World as both a training resource and a diagnostic benchmark for developing more robust and generalizable animal pose models in natural environments.

Limitations. Animal World currently focuses on image-level 2D pose estimation with ground-truth bounding boxes. It does not model temporal dynamics or provide manually verified 3D pose and shape ground truth, and its keypoint taxonomy primarily targets quadrupeds and primates. Future extensions could incorporate temporal annotations, broader anatomical coverage, end-to-end multi-animal perception, and real-world 3D supervision.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, improving clarity, and organizing parts of the manuscript and literature review. All scientific content, experimental results, analyses, citations, and claims were reviewed and verified by the authors. The authors take full responsibility for the final content of the paper.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

If authors feel that their paper submission raises questions regarding the Code of Ethics, they are encouraged to include a paragraph of Ethics Statement (at the end of the main text before references) to address potential concerns where appropriate. Topics include, but are not limited to, studies that involve human subjects, practices to data set releases, potentially harmful insights, methodologies and applications, potential conflicts of interest and sponsorship, discrimination/bias/fairness concerns, privacy and security issues, legal compliance, and research integrity issues (e.g., IRB, documentation, research ethics). This statement should not be more than 1 page.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

It is important that the work published in ICLR is reproducible. Authors are strongly encouraged to include a paragraph-long Reproducibility Statement at the end of the main text (before references) to discuss the efforts that have been made to ensure reproducibility. This paragraph should not itself describe details needed for reproducing the results, but rather reference the parts of the main paper, appendix, and supplemental materials that will help with reproducibility. For example, for novel models or algorithms, a link to an anonymous downloadable source code can be submitted as supplementary materials; for theoretical results, clear explanations of any assumptions and a complete proof of the claims can be included in the appendix; for any datasets used in the experiments, a complete description of the data processing steps can be provided in the supplementary materials. Each of the above are examples of things that can be referenced in the reproducibility statement.

AUTHOR CONTRIBUTIONS

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Muhammad Aamir, Naoya Muramatsu, Sangyun Shin, Matthew Wijers, Jia-Xing Zhong, Xinyu Hou, Amir Patel, Andrew Loveridge, and Andrew Markham. Wilddepth: A multimodal dataset for 3d wildlife perception and depth estimation. *arXiv preprint arXiv:2603.16816*, 2026.

Liang An, Jilong Ren, Tao Yu, Tang Hai, Yichang Jia, and Yebin Liu. Three-dimensional surface motion capture of multiple freely moving pigs using mammal. *Nature Communications*, 14(1): 7727, 2023.

Liang An, Jin Lyu, Li Lin, Pujin Cheng, Yebin Liu, and Xiaoying Tang. Animer+: Unified pose and shape estimation across mammalia and aves via family-aware transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3233–3249, 2026. doi: 10.1109/TPAMI.2025.3633828.

Mehmet Aygun and Oisin Mac Aodha. Saor: Single-view articulated object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10382–10391, 2024.

Marc Badger, Yufu Wang, Adarsh Modh, Ammon Perkes, Nikos Kolotouros, Bernd G Pfrommer, Marc F Schmidt, and Kostas Daniilidis. 3d bird reconstruction: a dataset, model, and shape recovery from a single view. In *European conference on computer vision*, pp. 1–17. Springer, 2020.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Praneet C Bala, Benjamin R Eisenreich, Seng Bum Michael Yoo, Benjamin Y Hayden, Hyun Soo Park, and Jan Zimmermann. Automated markerless pose estimation in freely moving macaques with openmonkeystudio. *Nature communications*, 11(1):4560, 2020.

Prianka Banik, Lin Li, and Xishuang Dong. A novel dataset for keypoint detection of quadruped animals from images, 2021. URL https://arxiv.org/abs/2108.13958.

Benjamin Biggs, Oliver Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out? 3d animal reconstruction with expectation maximization in the loop. In *European Conference on Computer Vision*, pp. 195–211. Springer, 2020.

Jinkun Cao, Hongyang Tang, Hao-Shu Fang, Xiaoyong Shen, Cewu Lu, and Yu-Wing Tai. Cross-domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9498–9507, 2019.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

Joo Young Choi, Wonkwang Lee, Ju-hyeong Seon, and Gunhee Kim. Every dog has its day, probably: A balanced synthetic benchmark and probabilistic modeling for 3d dog pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020.

Bo Hu, Bryan Seybold, Shan Yang, Avneesh Sud, Yi Liu, Karla Barron, Paulyn Cha, Marcelo Cosino, Ellie Karlsson, Janessa Kite, et al. 3d mouse pose from single-view video and a new dataset. *Scientific Reports*, 13(1):13554, 2023.

Xuyi Hu, Jin Lyu, Jiuming Liu, Yebin Liu, Silvia Zuffi, Liang An, and Stefan Goetz. Sam 3d animal: Promptable animal 3d reconstruction from images in the wild. *arXiv preprint arXiv:2605.07604*, 2026.

Chun-Hao P Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13274–13285, 2022.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023.

Bo Jin, Jin Lyu, Bin Zhang, Tao Yu, Liang An, Yebin Liu, Meili Wang, et al. Monocular mesh recovery and body measurement of female saanen goats. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 38670–38678, 2026.

Daniel Joska, Liam Clark, Naoya Muramatsu, Ricardo Jericevich, Fred Nicolls, Alexander Mathis, Mackenzie W Mathis, and Amir Patel. Acinoset: A 3d pose estimation dataset and baseline models for cheetahs in the wild. In *2021 IEEE international conference on robotics and automation (ICRA)*, pp. 13901–13908. IEEE, 2021.

Rollyn Labuguen, Jumpei Matsumoto, Salvador Blanco Negrete, Hiroshi Nishimaru, Hisao Nishijo, Masahiko Takada, Yasuhiro Go, Ken-ichi Inoue, and Tomohiro Shibata. Macaquepose: a novel "in the wild" macaque monkey pose dataset for markerless motion capture. *Frontiers in behavioral neuroscience*, 14:581154, 2021.

Chen Li and Gim Hee Lee. Coarse-to-fine animal pose and shape estimation. *Advances in Neural Information Processing Systems*, 34:11757–11768, 2021.

Ci Li, Ylva Mellbin, Johanna Krogager, Senya Polikovsky, Martin Holmberg, Nima Ghorbani, Michael J Black, Hedvig Kjellström, Silvia Zuffi, and Elin Hernlund. The poses for equine research dataset (pferd). *Scientific Data*, 11(1):497, 2024a.

Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10863–10872, 2019a.

Shuyuan Li, Jianguo Li, Hanlin Tang, Rui Qian, and Weiyao Lin. Atrw: A benchmark for amur tiger re-identification in the wild. *arXiv preprint arXiv:1906.05586*, 2019b.

Zizhang Li, Dor Litvak, Ruining Li, Yunzhi Zhang, Tomas Jakab, Christian Rupprecht, Shangzhe Wu, Andrea Vedaldi, and Jiajun Wu. Learning the 3d fauna of the web. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9752–9762, 2024b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Jin Lyu, Tianyi Zhu, Yi Gu, Li Lin, Pujin Cheng, Yebin Liu, Xiaoying Tang, and Liang An. Animer: Animal pose and shape estimation using family aware transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17486–17496, 2025.

Jin Lyu, Liang An, Pujin Cheng, Yebin Liu, and Xiaoying Tang. 4dequine: Disentangling motion and appearance for 4d equine reconstruction from monocular video. *arXiv preprint arXiv:2603.10125*, 2026.

Lucas Martini, Alexander Lappe, Anna Bognár, Rufin Vogels, and Martin A Giese. Bigmac3d: A big macaque motion and animation dataset bridging image and 3d pose representations. In *The Fourteenth International Conference on Learning Representations*.

Alexander Mathis, Thomas Biasi, Steffen Schneider, Mert Yuksekgonul, Byron Rogers, Matthias Bethge, and Mackenzie W Mathis. Pretraining boosts out-of-domain robustness for pose estimation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1859–1868, 2021.

Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19023–19034, 2022.

Tomasz Niewiadomski, Anastasios Yiannakidis, Hanz Cuevas-Velasquez, Soubhik Sanyal, Michael J Black, Silvia Zuffi, and Peter Kulits. Generative zoo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8492–8502, 2025.

Yichen Peng, Jyun-Ting Song, Chen-Chieh Liao, Kris Kitani, Hideki Koike, and Erwin Wu. Interpet4d: A multimodal 4d human-pet interaction dataset for pet motion generation. *arXiv preprint arXiv:2607.10287*, 2026.

Nadine Rüegg, Silvia Zuffi, Konrad Schindler, and Michael J Black. Barc: Learning to regress 3d dog shape from images by exploiting breed information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3876–3884, 2022.

Nadine Rüegg, Shashank Tripathi, Konrad Schindler, Michael J Black, and Silvia Zuffi. Bite: Beyond priors for improved three-d dog pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8867–8876, 2023.

Cristina Segalin, Jalani Williams, Tomomi Karigo, May Hui, Moriel Zelikowsky, Jennifer J Sun, Pietro Perona, David J Anderson, and Ann Kennedy. The mouse action recognition system (mars) software pipeline for automated analysis of social behaviors in mice. *Elife*, 10:e63720, 2021.

Moira Shooter, Charles Malleson, and Adrian Hilton. Benchmarking monocular 3d dog pose estimation using in-the-wild motion capture data. June 2024a.

Moira Shooter, Charles Malleson, and Adrian Hilton. Digidogs: Single-view 3d pose estimation of dogs using synthetic training data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 80–89, 2024b.

Samarth Sinha, Roman Shapovalov, Jeremy Reizenstein, Ignacio Rocco, Natalia Neverova, Andrea Vedaldi, and David Novotny. Common pets in 3d: Dynamic new-view synthesis of real-life deformable categories. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4881–4891. IEEE, 2023.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5693–5703, 2019.

Keqiang Sun, Dor Litvak, Yunzhi Zhang, Hongsheng Li, Jiajun Wu, and Shangzhe Wu. Ponymation: Learning articulated 3d animal motions from unlabeled online videos. In *European Conference on Computer Vision*, pp. 100–119. Springer, 2024.

Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 595–604, 2015.

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

Zan Wang, Siyu Chen, Luya Mo, Xinfeng Gao, Yuxin Shen, Lebin Ding, and Wei Liang. Dogmo: A large-scale multi-view rgb-d dataset for 4d canine motion recovery. *arXiv preprint arXiv:2510.24117*, 2025.

Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Magicpony: Learning articulated 3d animals in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8792–8802, 2023.

Jiacong Xu, Yi Zhang, Jiawei Peng, Wufei Ma, Artur Jesslen, Pengliang Ji, Qixin Hu, Jiehua Zhang, Qihao Liu, Jiahao Wang, et al. Animal3d: A comprehensive dataset of 3d animal pose and shape. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9099–9109, 2023.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022.

Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. Lasr: Learning articulated shape reconstruction from a monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15980–15989, 2021.

Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2863–2873, 2022a.

Yuxiang Yang, Junjie Yang, Yufei Xu, Jing Zhang, Long Lan, and Dacheng Tao. Apt-36k: A large-scale benchmark for animal pose estimation and tracking. *Advances in Neural Information Processing Systems*, 35:17301–17313, 2022b.

Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Lassie: Learning articulated shapes from sparse image ensemble via 3d part discovery. *Advances in Neural Information Processing Systems*, 35:15296–15308, 2022.

Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. *arXiv preprint arXiv:2108.12617*, 2021.

Xiaohang Yu, Ti Wang, and Mackenzie Weygandt Mathis. Prima: Boosting animal mesh recovery with biological priors and test-time adaptation, 2026. URL https://arxiv.org/abs/2606.02366.

Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. Humbi: A large multiview dataset of human body expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2990–3000, 2020.

Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution vision transformer for dense predict. *Advances in neural information processing systems*, 34:7281–7293, 2021.

Jakub Zadrozny, Oisin Mac Aodha, and Hakan Bilen. Bootstrapping articulated 3d reconstruction from 2d image collections. In *European Conference on Computer Vision*, pp. 557–574. Springer, 2026.

Brian Nlong Zhao, Jiajun Wu, and Shangzhe Wu. Web-scale collection of video data for 4d animal reconstruction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Silvia Zuffi and Michael J Black. Awol: Analysis without synthesis using language. In *European Conference on Computer Vision*, pp. 1–19. Springer, 2024.

Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6365–6373, 2017.

Silvia Zuffi, Angjoo Kanazawa, and Michael J Black. Lions and tigers and bears: Capturing non-rigid, 3d, articulated shape from images. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 3955–3963, 2018.

A Appendix

A.1 Cross Species Generalization

Table 9 evaluates cross-family generalization by excluding each target family from training. Performance consistently decreases on the held-out family, revealing a clear generalization gap to unseen animal categories. The magnitude of this drop varies across families, with larger degradation for Felidae and Equidae in our setting. This trend is consistent with prior observations on AP-10K and APT-36K that cross-family transfer depends strongly on the morphological and taxonomic similarity between seen and unseen animals Yu et al. (2021); Yang et al. (2022b).

Table 9: **AP_{50} Results of ViTPose-Base models for cross species evaluation of Animal World.**

Variant	Canidae	Felidae	Bovidae	Equidae
w/o Canidae	73.0	77.5	81.0	90.2
w/o Felidae	77.6	67.2	80.6	88.8
w/o Bovidae	77.5	77.2	76.8	91.0
w/o Equidae	78.2	77.5	81.7	79.9

A.2 Cross-Dataset Transfer on Shared Keypoints

Table 10 evaluates cross-dataset transfer using the 16 semantically shared keypoints. Both models exhibit clear performance degradation when transferred across datasets, confirming a substantial domain gap between APTv2 and Animal World. Notably, the Animal World-trained model transfers more effectively to APTv2 (59.6 AP) than the APTv2-trained model transfers to Animal World (47.2 AP). This asymmetric transfer suggests that the broader morphological and scene diversity in Animal World provides supervision that generalizes more effectively beyond its source domain, although a considerable gap to in-domain performance remains.

A.3 Long Tail Experiment

Animal World exhibits a long-tailed training distribution over species folders (counting annotated instances; multi-species clips A_B are treated as a single class). A natural concern is whether pose accuracy degrades for low-frequency taxa. To test this, we rank species by their training instance counts and partition the validation set into *Head*/*Mid*/*Tail* bins with approximately equal validation mass ($\sim$620 instances each), so that each bin is statistically comparable. In parallel, we compute per-instance OKS and measure its correlation with $\log_{10}(1 + n_{\text{train}})$. As shown in Table 11, AP_{50} is essentially flat across bins, and Tail AP (79.3) matches Mid (80.6); instance OKS is only weakly related to training frequency (Pearson $r{=}-0.07$). Thus, although the label distribution is long-tailed, we do not observe a classic long-tail performance collapse on rare species. The modest drop on Head is concentrated on a few high-frequency, visually challenging taxa rather than a frequency–accuracy trend.

A.4 Leakage Control

To quantify train–validation leakage under video-derived wildlife imagery, we perform a perceptual near-duplicate analysis. For every validation image, we extract a 64-bit pHash and compute the minimum Hamming distance to all training images, converting distance d to similarity $(64 - d)/64$. Following common practice for 64-bit pHash, we treat $d \leq 5$ ($\geq$92.2% similarity) as near-duplicate leakage and $6 \leq d \leq 10$ as borderline. Table 12 refers that, on the Animal World train/val split, no validation image falls into either regime: all 1004 validation images have minimum same-species Hamming distance $d \geq 11$ ($\leq$82.8% similarity), indicating no measurable near-duplicate leakage.

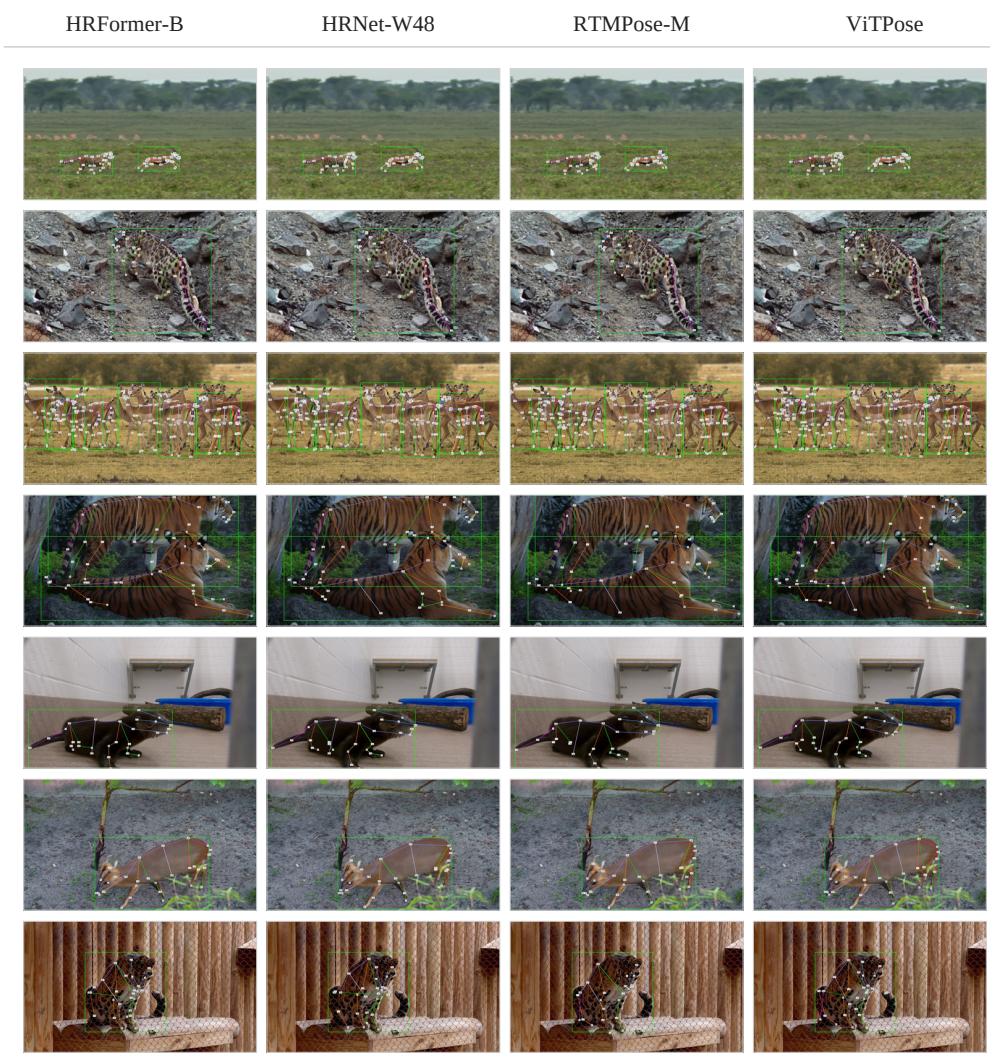
A.5 Baseline Results

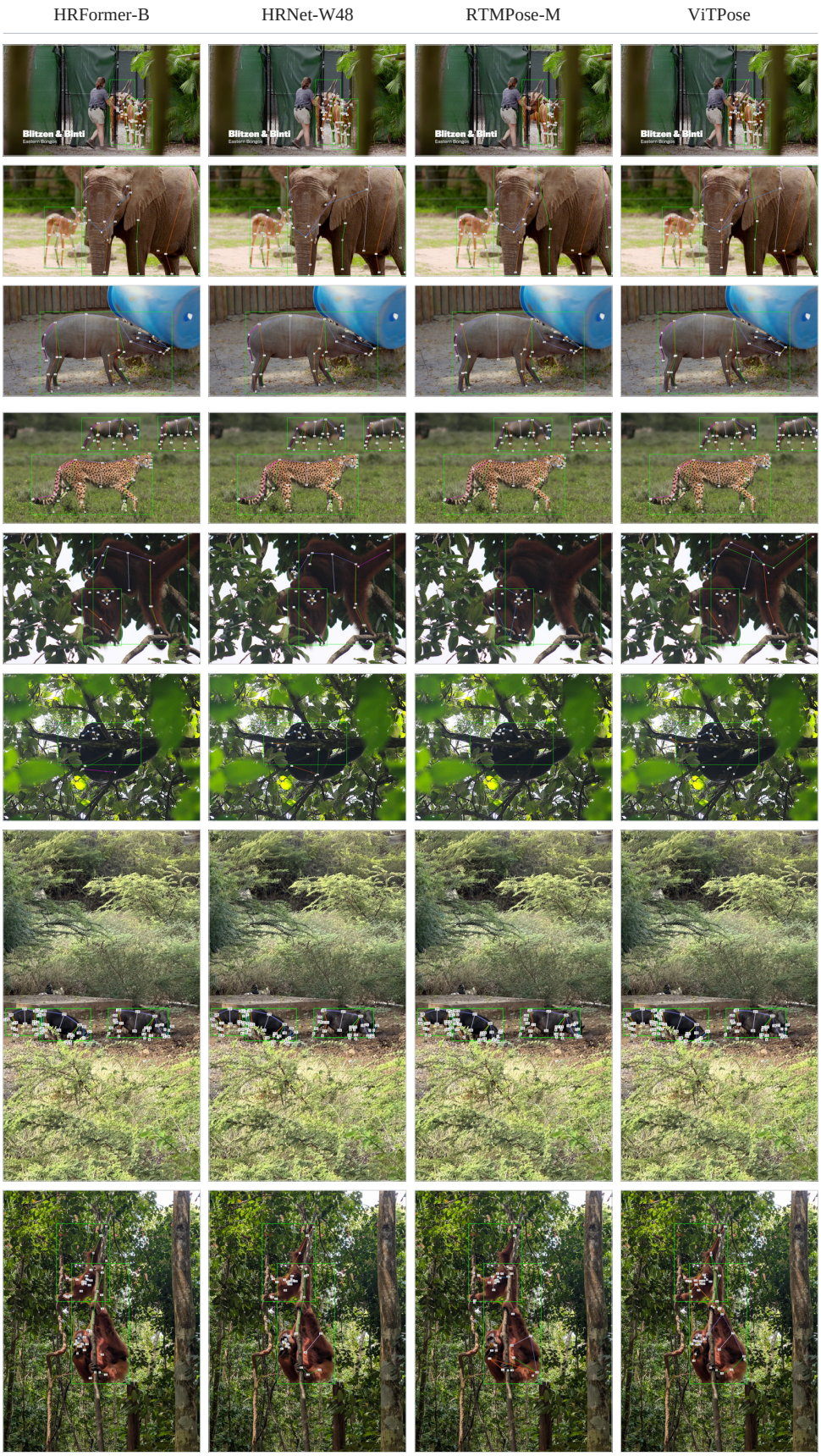
Table 10: **Cross-dataset transfer on shared keypoints.** ViTPose-B is trained on APTv2 and Animal World and evaluated on both datasets using the 16 semantically matched keypoints.

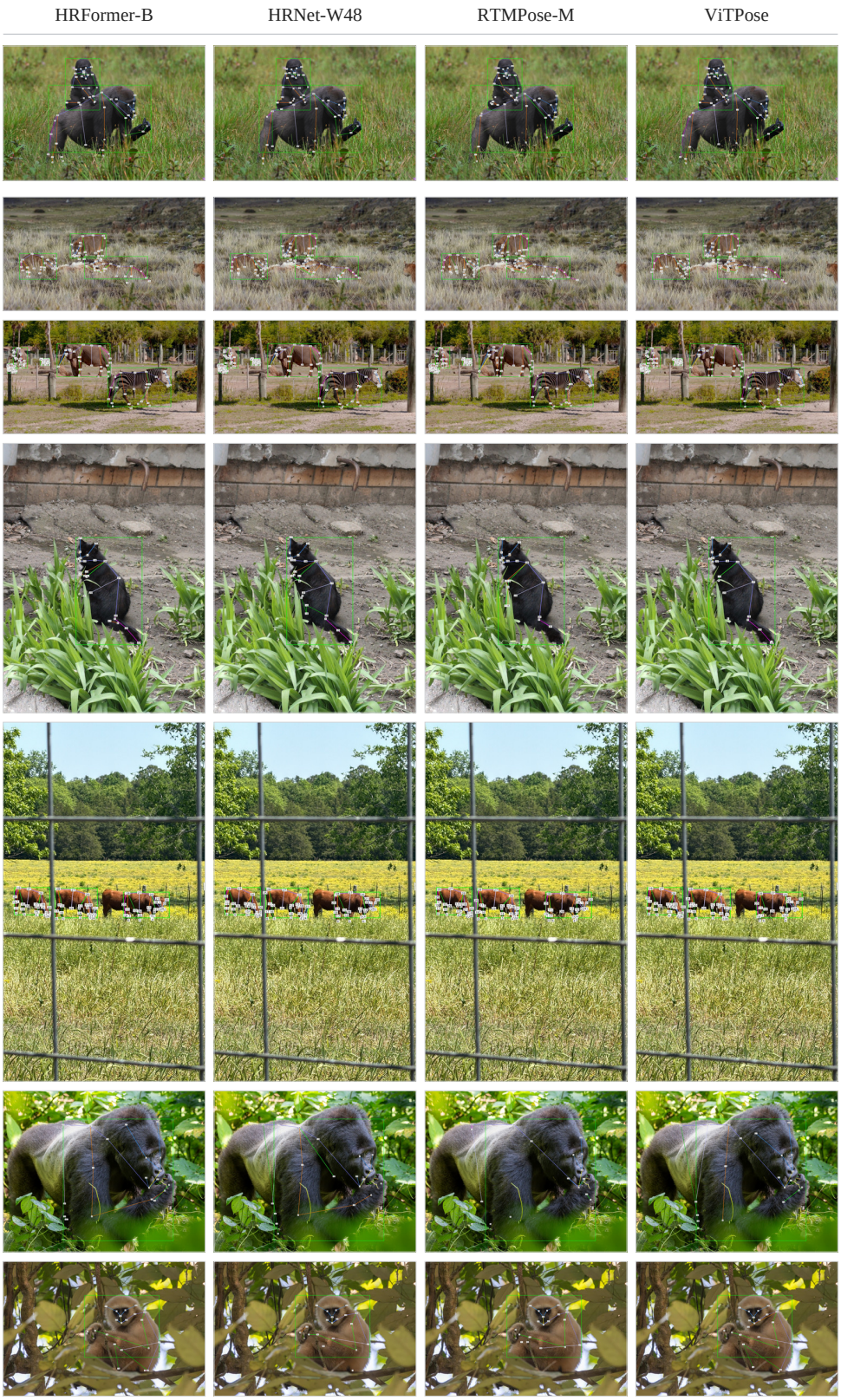
Train	APTv2 Test		Animal World Test	
	AP_{50}	AP	AP_{50}	AP
APTv2	**92.4**	**76.5**	89.8	47.2
Animal World	88.1	59.6	**96.9**	**80.1**

Table 11: **Long-tail analysis by training species frequency.** Species folders are ranked by train instance counts and partitioned so that each bin contains roughly one-third of the validation instances. We report COCO AP_{50}/mAP and mean instance OKS. Despite a long-tailed training distribution, low-frequency (*Tail*) species do not underperform relative to mid-frequency species.

Bin	Images	Inst.	AP_{50} ↑	AP ↑	OKS ↑
Head (train $n \geq 1049$)	294	597	95.9	74.8	0.860
Mid (215–850)	340	634	96.8	80.6	0.892
Tail (train $n \leq 214$)	367	634	96.9	79.3	0.885




HRFormer-B
HRNet-W48
RTMPose-M
ViTPose


HRFormer-B
HRNet-W48
RTMPose-M
ViTPose

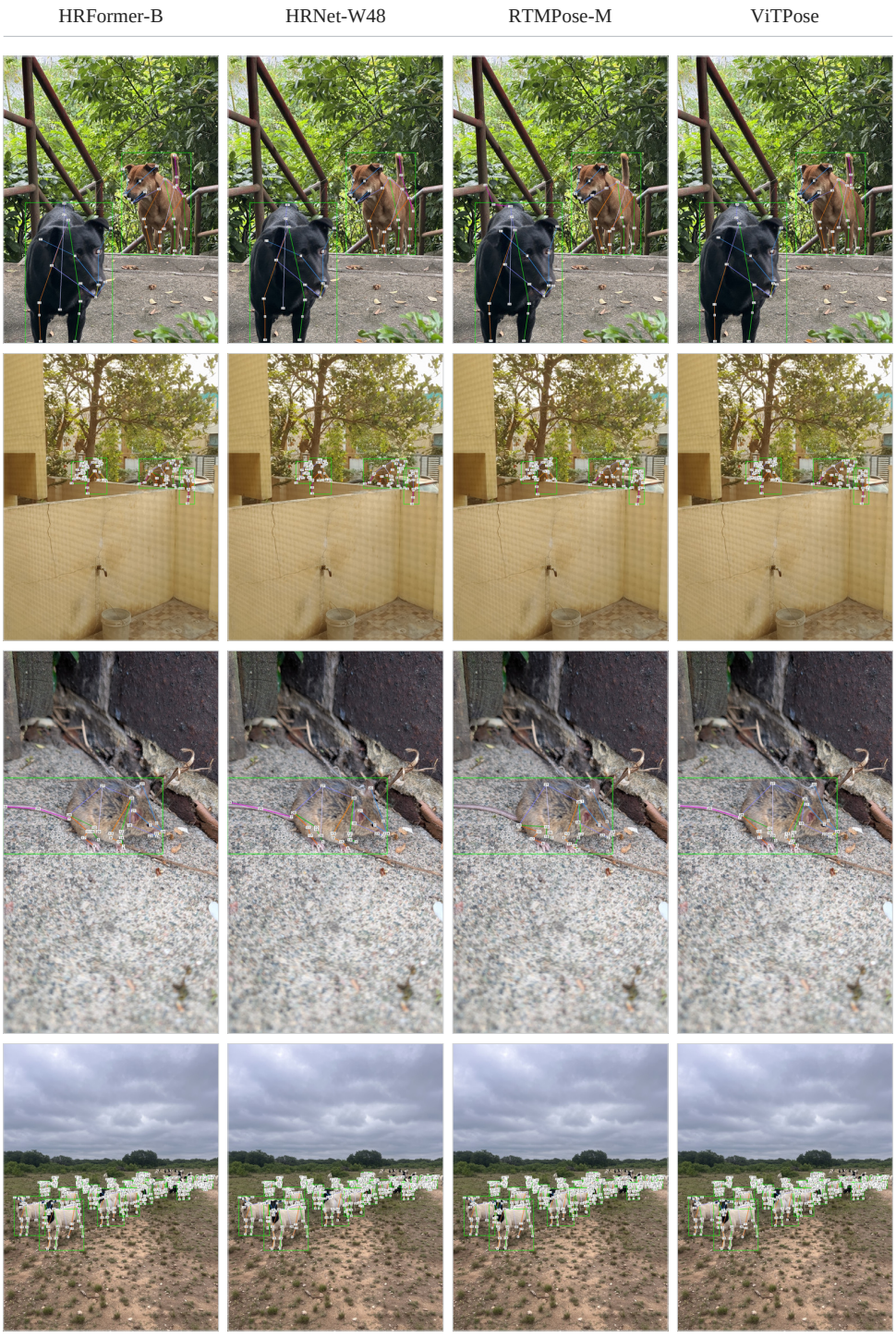


Figure 6: Qualitative results on the Animal World Dataset. Each row shows one input frame; columns compare HRFormer-B, HRNet-W48, RTMPose-M, and ViTPose.

A.6 Dataset Statistics

Table 12: **Near-duplicate analysis after train/val de-duplication.** For each validation image, we compute a 64-bit pHash and report the minimum Hamming distance to training images. Similarity is defined as $(64 - d)/64$.

Min Hamming	Similarity	Count	%	Interpretation
0–5	92.2–100%	0	0.0	Near-duplicate (leakage)
6–10	84.4–90.6%	0	0.0	Borderline
$\geq$11	$\leq$82.8%	1004	100.0	Unlikely leakage

Table 13: **Per-folder statistics of Animal World.** Each row corresponds to a top-level folder under the dataset root. Cross-species scenes are written with "&" between species names. We report the number of species encoded in the folder name, images, and animal instances. Cross-species folders are not counted as additional species, and "internet" variants (e.g., cat internet) are not treated as separate species from their non-internet counterparts. Rare follows our IUCN CR/EN split (a folder is marked rare if any constituent species is Critically Endangered or Endangered). Age indicates whether the folder contains age-diversity (baby/family) annotations. Totals: 221 folders (124 single-species, 97 cross-species; 46 rare, 74 with age), 10036 images, 19132 instances. Rows are sorted alphabetically by dataset folder name.

Species / combination	Species	Cross-species	Rare	Age	Images	Instances
african buffalo	1	N	N	Y	163	402
african buffalo & elephant	2	Y	Y	N	2	7
african buffalo & lion	2	Y	N	N	53	188
african wildcat	1	N	N	N	9	11
african wilddog	1	N	Y	Y	246	561
african wilddog & elephant	2	Y	Y	N	4	18
african wilddog & greater kudu	2	Y	Y	N	7	23
african wilddog & lechwe	2	Y	Y	N	3	7
ankole watusi cattle	1	N	N	N	2	3
ankole watusi cattle & zebra internet	2	Y	N	N	1	3
arctic wolf	1	N	N	N	7	16
babirusa	1	N	N	N	6	6
baboon	1	N	N	Y	119	246
baboon & elephant	2	Y	Y	N	4	20
baboon & greater kudu	2	Y	N	Y	2	11
baboon & hyena	2	Y	N	N	1	3
baboon & impala	2	Y	N	N	8	28
baboon & impala & elephant	3	Y	Y	N	1	15
baboon & impala & zebra	3	Y	N	N	1	6
badger	1	N	N	N	5	7
bighorn sheep	1	N	N	Y	17	46
bison	1	N	N	Y	28	53
black bear	1	N	N	Y	35	52
black bear & brown bear	2	Y	N	N	5	20
black bear & wild boar	2	Y	N	N	1	5
bongo calf	1	N	N	N	7	11
bornean orangutan internet	1	N	Y	Y	44	52
brown bear	1	N	N	Y	227	384
bush buck	1	N	N	N	1	1
camel	1	N	N	N	12	34
capybara	1	N	N	Y	5	14
caracal	1	N	N	N	18	18
cat	1	N	N	N	10	21
cat internet	1	N	N	N	169	172
cattle	1	N	N	Y	17	44
cattle & dog	2	Y	N	N	1	3
cattle & wolf	2	Y	N	N	2	11
cattle & zebra internet	2	Y	N	N	4	21
chamois	1	N	N	N	2	3
cheetah	1	N	N	Y	534	873

Continued on next page

Table 13: (continued) Per-folder statistics of Animal World.

Species / combination	Species	Cross-species	Rare	Age	Images	Instances
cheetah & gazelle	2	Y	N	N	20	47
cheetah & gazelle & wildebeest	3	Y	N	N	1	3
cheetah & impala	2	Y	N	N	11	31
cheetah & jackal	2	Y	N	N	1	3
cheetah & sable antelope	2	Y	N	N	5	10
cheetah & topi	2	Y	N	N	3	10
cheetah & wildebeest	2	Y	N	N	25	77
chimpanzee	1	N	Y	Y	7	13
chimpanzee internet	1	N	Y	Y	39	46
civet	1	N	N	N	6	7
clouded leopard	1	N	N	N	12	12
cougar	1	N	N	Y	17	28
cow internet	1	N	N	Y	300	558
coyote	1	N	N	N	2	2
deer	1	N	N	Y	61	105
dik dik	1	N	N	N	4	4
dog internet	1	N	N	Y	251	276
duiker	1	N	N	Y	8	10
eastern gorilla internet	1	N	Y	Y	29	38
eland	1	N	N	Y	17	47
elephant	1	N	Y	Y	1070	2436
elephant & baboons	2	Y	Y	Y	5	19
elephant & gazelle	2	Y	Y	N	1	4
elephant & gazelle & zebra	3	Y	Y	N	4	23
elephant & giraffe & oryx	3	Y	Y	N	3	16
elephant & hippo	2	Y	Y	N	2	7
elephant & impala	2	Y	Y	Y	9	36
elephant & impala & african buffalo	3	Y	Y	N	2	19
elephant & lion	2	Y	Y	Y	35	113
elephant & oryx	2	Y	Y	N	1	2
elephant & rhinoceros	2	Y	Y	N	1	2
elephant & wild boar	2	Y	Y	N	1	3
elephant & wildebeest	2	Y	Y	N	2	10
elephant & zebra	2	Y	Y	N	27	109
fallow deer	1	N	N	N	8	15
fox	1	N	N	Y	100	151
gazelle	1	N	N	Y	105	206
gazelle & jackal	2	Y	N	N	14	33
gazelle & mongoose	2	Y	N	N	1	4
gazelle & wildebeest	2	Y	N	N	1	2
gelada	1	N	N	Y	2	8
gibbon	1	N	Y	N	1	1
gibbons internet	1	N	Y	Y	24	27
giraffe	1	N	N	Y	287	701
giraffe & elephant	2	Y	Y	N	4	15
giraffe & hyena	2	Y	N	N	6	19
giraffe & lion	2	Y	N	N	8	22
giraffe & rhinoceros	2	Y	Y	N	2	4
giraffe & springbok	2	Y	N	N	1	2
giraffe & wild boar	2	Y	N	N	1	7
giraffe & wildebeest	2	Y	N	N	8	62
giraffe & zebra	2	Y	N	N	12	68
giraffe & zebra & waterbuck	3	Y	N	N	1	4
goat	1	N	N	N	8	18
gorilla	1	N	Y	Y	6	11
great pyrenees	1	N	N	N	8	23
greater kudu	1	N	N	Y	11	25
guanaco	1	N	N	N	11	50
hare	1	N	N	N	18	34
herding dog	1	N	N	N	1	1

Continued on next page

Table 13: (continued) Per-folder statistics of Animal World.

Species / combination	Species	Cross-species	Rare	Age	Images	Instances
hippo	1	N	N	Y	71	122
hippo & african wilddog	2	Y	Y	N	1	3
horse	1	N	N	N	4	11
hyena	1	N	N	Y	654	1162
hyena & african wilddog	2	Y	Y	N	4	9
hyena & cheetah	2	Y	N	N	5	12
hyena & cheetah & jackal	3	Y	N	N	1	3
hyena & gazelle	2	Y	N	N	4	14
hyena & impala	2	Y	N	N	6	20
hyena & jackal	2	Y	N	N	10	24
hyena & jackal & wildebeest	3	Y	N	N	3	10
hyena & topi	2	Y	N	N	3	6
hyena & topi & gazelle	3	Y	N	N	1	2
hyena & wildebeest	2	Y	N	N	2	5
hyena & zebra	2	Y	N	N	2	4
iberian ibex	1	N	N	Y	46	82
ibex	1	N	N	Y	12	33
impala	1	N	N	Y	311	762
impala & baboons	2	Y	N	N	4	19
impala & zebra	2	Y	N	N	15	66
impala & zebra & elephant	3	Y	Y	N	1	4
indian bison	1	N	N	N	1	2
jackal	1	N	N	Y	159	234
jaguar	1	N	N	N	99	107
kangaroo	1	N	N	Y	47	93
lechwe	1	N	N	Y	22	38
lemur	1	N	Y	N	3	3
leopard	1	N	N	Y	343	371
leopard & african buffalo	2	Y	N	N	1	2
leopard & hippo	2	Y	N	N	1	3
leopard & hyena	2	Y	N	N	2	4
leopard & impala	2	Y	N	N	1	5
leopard & zebra & waterbuck	3	Y	N	N	1	7
lion	1	N	N	Y	1307	2396
lion & hippo	2	Y	N	N	14	40
lion & hyena	2	Y	N	N	21	67
lion & hyena & jackal	3	Y	N	N	1	3
lion & impala	2	Y	N	N	4	14
lion & jackal	2	Y	N	N	2	6
lion & oryx	2	Y	N	N	1	3
lion & rhinoceros	2	Y	Y	N	3	18
lion & springbok	2	Y	N	N	3	13
lion & topi	2	Y	N	N	1	3
lion & warthog	2	Y	N	N	1	2
lion & waterbuck	2	Y	N	N	1	4
lion & wildebeest	2	Y	N	N	9	23
lion & wildebeest & zebra	3	Y	N	N	7	25
lion & zebra	2	Y	N	N	9	19
lynx	1	N	N	Y	52	64
mandrills	1	N	N	N	1	2
margay	1	N	N	N	1	1
markhor	1	N	N	N	2	2
marmot	1	N	N	N	2	4
meerkat	1	N	N	Y	22	58
mongolian gazelle	1	N	N	N	1	2
mongoose	1	N	N	N	11	15
monkey	1	N	N	Y	21	30
monkey internet	1	N	N	Y	156	210
moose	1	N	N	Y	61	93
mouse internet	1	N	N	N	84	84

Continued on next page

Table 13: (continued) Per-folder statistics of Animal World.

Species / combination	Species	Cross-species	Rare	Age	Images	Instances
musk oxen	1	N	N	N	1	5
nyala	1	N	N	Y	36	61
okapi	1	N	Y	N	2	2
opossum	1	N	N	N	4	5
oribi	1	N	N	N	1	1
oryx	1	N	N	N	19	29
otter	1	N	N	N	3	3
pampas deer	1	N	N	N	12	23
pig	1	N	N	Y	3	6
pig internet	1	N	N	Y	92	156
polar bear	1	N	N	Y	95	123
pronghorn	1	N	N	Y	23	55
puku	1	N	N	Y	41	82
rabbit	1	N	N	N	1	2
rat	1	N	N	N	2	2
red deer	1	N	N	Y	81	211
reindeer	1	N	N	Y	19	79
rhinoceros	1	N	Y	Y	166	256
roe deer	1	N	N	Y	15	18
sable antelope	1	N	N	N	1	2
saiga antelope	1	N	Y	Y	26	46
serval	1	N	N	Y	36	40
sheep internet	1	N	N	Y	154	331
shunk	1	N	N	N	3	3
sifaka	1	N	Y	N	2	2
sloth bear	1	N	N	N	1	1
snow leopard	1	N	N	N	27	33
springbok	1	N	N	Y	31	105
springbok & zebra	2	Y	N	N	1	5
squirrel	1	N	N	N	10	10
sumatran orangutan internet	1	N	Y	Y	49	59
tibetan antelope	1	N	N	Y	7	20
tibetan fox	1	N	N	Y	7	9
tiger	1	N	Y	Y	112	125
topi	1	N	N	Y	36	81
warthog	1	N	N	Y	58	92
waterbuck	1	N	N	N	4	7
western gorilla & deer internet	2	Y	Y	N	1	3
western gorilla internet	1	N	Y	Y	136	183
white fox	1	N	N	Y	12	17
white tiger	1	N	Y	Y	16	41
wild boar	1	N	N	Y	54	98
wild goat	1	N	N	N	3	4
wildebeest	1	N	N	Y	154	364
wildebeest & african wilddog	2	Y	Y	N	1	2
wildebeest & hippo & zebra	3	Y	N	N	2	10
wolf	1	N	N	Y	234	405
wolf & bison	2	Y	N	N	2	10
wolf & dog	2	Y	N	N	1	2
wolverine	1	N	N	N	30	32
wombat	1	N	N	N	2	2
zebra	1	N	N	Y	136	287
zebra & african wilddog	2	Y	Y	N	1	4
zebra & eland	2	Y	N	N	2	8
zebra & giraffe & wildebeest	3	Y	N	N	1	4
zebra & jackal	2	Y	N	N	1	3
zebra & jackal & springbok	3	Y	N	N	1	3
zebra & wildebeest	2	Y	N	N	10	36
zebra & wildebeest & springbok	3	Y	N	N	1	6
zebra & wildebeest & topis	3	Y	N	N	1	5

Continued on next page

Table 13: (continued) Per-folder statistics of Animal World.

Species / combination	Species	Cross-species	Rare	Age	Images	Instances
zebu cattle	1	N	N	N	1	12

Total: 221 folders, 10036 images, 19132 instances.