Generative Zoo

Supplementary Material

A. Training Details

Our model is trained with three losses: a joint 2D-projection L1 with weight `0.01`, an 9D-rotation-matrix MSE after performing symmetric orthogonalization with weights of `100` on `body_pose` and `global_orient` (following ablations performed in Geist et al. [2]), and an L1 loss on transformed vertices after applying `betas` with a weight of `50`. A batch size of `128` with a single GPU is used across experiments. We configure early stopping based on validation joint 2D-projection loss.

B. Perceptual Study

Motivated by inconsistencies observed during evaluation on Animal3D [6] (see Sec. 4.1 and Fig. 7), we perform a perceptual study comparing our predictions against the Animal3D ground truth to investigate our hypothesis that there is an upper limit on achievable quantitative performance. We show 48 participants on Amazon Mechanical Turk (AMT) a set of 22 randomly selected dataset samples along with five warm-up samples and three catch trials. Each sample consists of the source image and side renders of both the ground-truth and predicted meshes. Participants are tasked with determining which of the two meshes is posed in a way that is better aligned with the animal in the image. Warm-up samples are discarded prior to analysis, and the five participants that failed two or more catch trials are excluded. However, we note that quantities reported below do not change when participants are not excluded.

For each of the samples, we perform a one-sided binomial test to determine whether the predicted mesh is preferred over the ground truth mesh. We find that the predicted mesh is *significantly* preferred at $\alpha = 0.05$ in 27% of the samples. To correct for multiple comparisons, we apply the Benjamini–Hochberg correction across tests, and find that the result remains significant. In response, we reject the null hypothesis that ground-truth samples are consistently preferable.

C. Image-Generation Model Ablation

We additionally ablate our choice of FLUX as the image-generation model. We compare against Hunyuan-DiT [4] and Stable Diffusion 3 [1]. Training on 100,000 samples for each experiment, we observe greatest performance training on images produced using FLUX. We report quantitative results in Tab. 1 and include a visual comparison in Fig. 1. While PCK scores remain somewhat saturated across experiments, as observed during earlier experiments, we observe more-notable differences in 3D metrics, with FLUX outperforming Hunyuan-DiT and Stable Diffusion 3 placing third. We observe qualitatively that the Stable Diffusion 3 model can struggle producing outputs aligned with the control signal, and that both of the non-FLUX models produce comparatively unrealistic images. Even without control signals, we find that Hunyuan-DiT produces samples that appear cartoon-like, while Stable Diffusion 3 generations are made more natural.

Figure 1. **Image-Generation-Model Ablation.** We ablate our choice of FLUX [3] as the image-generation model, comparing against Hunyuan-DiT [4] and Stable Diffusion 3 (SD3) [1]. We observe that both ablated models produce results that are less visually realistic than those of FLUX. Hunyuan-DiT appears to produce more cartoon-like samples, whereas we observe more-frequent control-signal failures with Stable Diffusion 3.

D. Evaluating on GenZoo-Felidae

To help assess the realism of our synthetic data, we evaluate a model trained only on Animal3D [6] on *GenZoo-Felidae*. We observe comparable 2D but worse 3D metrics (Tab. 2), supporting its realism, but also highlighting the difficulty of producing accurate 3D labels in the wild.

E. Occlusion Augmentation

As we note in Sec. 5 and Fig. 11, we observe that our model is not robust to strong occlusions. We evaluate whether this can be improved through train-time data augmentation. Following the approach of Sárándi et al. [5], we apply occlusion augmentations in the form of Pascal VOC objects (excluding animal-object augmentations) with a 10% probability. We observe improvements across metrics (see Tab. 3), suggesting its effectiveness.

References

[1] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow

	Animal3D			GenZoo-Felidae				
	↑ PCK@0.5	↓ S-MPJPE	↓ PA-MPJPE	↑ PCK@0.5	↓ S-MPJPE	↓ PA-MPJPE	↓ S-V2V	↓ PA-V2V
FLUX [3]	<u>97.1</u>	**166.9**	**118.4**	**99.6**	**83.5**	**62.0**	**91.5**	**72.1**
H-DiT [4]	95.9	<u>174.0</u>	<u>125.6</u>	99.0	<u>99.9</u>	<u>74.4</u>	<u>106.9</u>	<u>83.3</u>
SD3 [1]	**97.5**	178.3	127.8	<u>99.3</u>	109.0	81.8	122.4	97.3

Table 1. **Image-Generation-Model Ablation Effects**. We ablate our choice of image-generation model, generating datasets of 100k samples each. We observe that FLUX outperforms the ablated models, Hunyuan-DiT and Stable Diffusion 3, in the majority of metrics.

	↑ PCK@0.5	↓ S-MPJPE	↓ PA-MPJPE	↓ S-V2V	↓ PA-V2V
Ours	**98.6**	**74.8**	**54.7**	**82.3**	**64.1**
Animal3D	<u>98.0</u>	<u>117.8</u>	<u>76.9</u>	<u>144.4</u>	<u>97.0</u>
Animal3D (ResNet)	92.7	174.5	107.0	190.6	121.8

Table 2. **Real2Sim**. We evaluate a model trained exclusively on Animal3D [6] on our *GenZoo-Felidae* to validate transferability.

	↑ PCK@0.5	↓ S-MPJPE	↓ PA-MPJPE
Occlusions	**98.1**	**155.0**	**115.2**
No occlusions	97.1	160.1	116.6

Table 3. **Occlusion augmentation**. We train our model on *GenZoo* 1M dataset, with 10% chance of occlusion augmentation. When evaluating on Animal3D [6], we observe improved performance of the model, demonstrating effectiveness of the occlusion augmentation during training.

Figure 2. **Results Beyond Animal3D.** Further qualitative reconstruction samples of the model trained on our synthetic data.

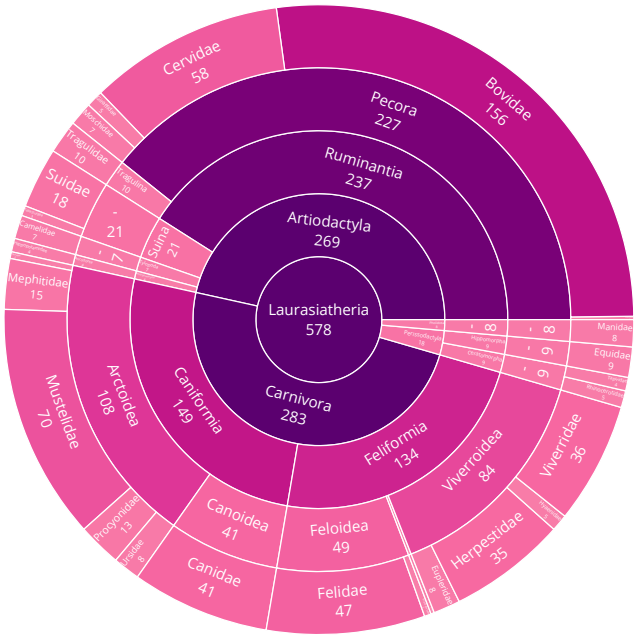

Figure 3. **Taxonomy.** We sample species from a subset of the mammalian Superclass Laurasiatheria. The figure displays the abbreviated taxonomical hierarchy of our sampling, where hyphens represent an empty level and the numbers are of contained species.

transformers for high-resolution image synthesis. In *ICML*, pages 12606–12633. PMLR, 2024. 1, 2

[2] Andreas René Geist, Jonas Frey, Mikel Zhobro, Anna Levina, and Georg Martius. Learning with 3D rotations, a hitchhiker's guide to SO(3). In *ICML*, pages 15331–15350. PMLR, 2024. 1

[3] Black Forest Labs. FLUX. https://github.com/black-forest-labs/flux, 2022. 1, 2

[4] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu.

Hunyuan-DiT: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024. 1, 2

[5] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe. How robust is 3D human pose estimation to occlusion? In *IROS Workshops*, 2018. 1

[6] 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 *ICCV*, pages 9099–9109, 2023. 1, 2