TaxaDiffusion: Progressively Trained Diffusion Model for Fine-Grained Species Generation

Amin Karimi Monsefi [1] Mridul Khurana [2] Rajiv Ramnath [1] Anuj Karpatne [2]
Wei-Lun Chao [1] Cheng Zhang [3]

[1] The Ohio State University [2] Virginia Tech [3] Texas A&M University

Contents

A. Additional Implementation Details

We provide implementation details omitted in the main text. Our model and code will be publicly available.

A.1. Hyperparameters

Table S1 summarizes the hyperparameters used across the models, including the baselines and our proposed TaxaDiffusion. All methods use the Stable Diffusion v1.5 [7] as the base generator. We employ consistent settings, such as the image resolution (512 × 512) and latent space dimensions (z-shape: $64 \times 64 \times 4$), to ensure comparability.

Unlike full fine-tuning, which requires updating all 859.52 million parameters of the model, our method, TaxaDiffusion, trains only a subset of parameters — accounting for just 6.5% of the total model size. This approach, utilizing 55.93 million trainable parameters, represents a significantly smaller fraction compared to full fine-tuning. Despite this reduction, our progressive taxonomy — based training strategy amplifies the impact of these updates by leveraging hierarchical information from higher taxonomic levels. This enables our method to achieve superior generative performance with a considerably smaller training footprint, highlighting the efficiency and effectiveness of our targeted approach.

We use a guidance scale of 6 across all models during inference, balancing image fidelity and trait specificity (see Section B.5 for more details). We set diffusion sampling step to 250 for consistent evaluation across all approaches.

We use cross-attention for all models as the conditioning mechanism and a batch size of 12 per GPU and a total batch size of 192. We use a learning rate of 1×10^{-4} for *SD + LoRA* and 1×10^{-5} for *SD + Full Finetuning*. For our method TaxaDiffusion, we use a learning rate of 1×10^{-4} for the first step because of having LoRA, and then 1×10^{-5} for the rest of the steps.

A.2. Training Details

We leverage a progressive taxonomy-based approach, where the model learns hierarchical information from higher taxonomic levels, such as `Family` and `Order`, before specializing in fine-grained traits at the `Species` level. This progressive strategy ensures that shared characteristics or traits at higher levels are effectively utilized, providing a strong foundation for generating accurate and semantically aligned species-specific images. This approach significantly enhances the model's ability to capture fine-grained details without the need for full model fine-tuning.

To efficiently train the model, we employ LoRA, which introduces low-rank (4 in all experiments) updates targeted specifically at key attention layers. These updates focus on critical components of the cross-attention mechanism, such as the queries, keys, and values. The LoRA configuration balances the rank of the updates and their scaling factor to ensure they effectively influence the model during training. By using LoRA and our hierarchical conditioning, we only train 55.93 million parameters — approximately 6.5% of the total model size — compared to the 859.52 million parameters updated during full fine-tuning. This efficiency enables us to achieve results comparable to full fine-tuning

Table S1. **Hyperparameter setting**. Training and inference stage hyperparameter details of all the baselines and TaxaDiffusion. †: We use a learning rate of 1×10^{-4} only for training the first level of TaxaDiffusion and the LoRA stage. *: TaxaDiffusion steps are mentioned for each level training. SD: Stable Diffusion 1.5.

Parameter	SD	SD + LoRA	SD + Full Finetuning	TaxaDiffusion
Image resolution	512×512	512×512	512×512	512×512
z-shape	$64 \times 64 \times 4$	$64 \times 64 \times 4$	$64 \times 64 \times 4$	$64 \times 64 \times 4$
Model Size	859.52 M	860.32 M	859.52 M	915.45 M
Trainable Parameters	0	0.80 M	859.52 M	55.93 M
Batch Size per GPU	–	12	12	12
Batch Size	–	192	192	192
Iterations	–	250K Steps	100K Steps	250K Steps*
Learning Rate	–	1×10^{-4}	1×10^{-5}	1×10^{-5}†
Guidance Scale	6	6	6	6

while significantly reducing computational overhead.

B. Additional Results and Analyses

B.1. Detailed Results on More Spices

The results in Table S2 demonstrate that classes with a higher number of diverse species at the `Class` taxonomic level achieve better results. This can be attributed to our hierarchical training approach and the integration of TaxaGuide. By progressively training from higher taxonomy levels, such as `Family` and `Order`, to lower levels, such as `Genus` and `Species`, our method effectively transfers shared traits and structural information across related species. This progressive strategy enhances the model's ability to generate detailed and accurate images for classes with richer inter-species diversity.

High-Species Diversity Classes. Some classes, such as “Actinopteri” and “Elasmobranchii”, include 400 species, achieved the best FID and BioCLIP scores (27.75 and 13.78 for “Actinopteri”). These results highlight the model's capability to handle diverse species datasets effectively. The relatively low LPIPS values for these classes (0.7236 and 0.7324) indicate high perceptual similarity between generated and real images, suggesting that a large number of species provides sufficient training diversity.

Low-Species-Diversity Classes. Some classes with fewer species (“Dipneusti”, “Myxini”, and “Cladistii”) experienced significantly higher FID values, such as 159.77 for “Dipneusti”. Despite generating 50 images per species to balance these classes, the limited inter-species diversity likely constrains the ability of the model to generalize, resulting in reduced visual fidelity. Similarly, lower BioCLIP scores in these classes (8.87 for *Dipneusti*) indicate weaker semantic alignment, reflecting challenges in capturing fine-grained distinctions when species representation is sparse.

Intermediate Classes. Some classes such as “Holocephali” and “Petromyzonti” with moderate species diversity (29 and 24 species, respectively) showed better FID and BioCLIP scores than “Dipneusti” but lagged behind “Actinopteri” and “Elasmobranchii”. The FID scores (109.48 and 93.14) indicate reasonable generation fidelity, while the BioCLIP scores (13.10 and 9.91) suggest effective, though not perfect, semantic alignment. This trend underscores the importance of sufficient inter-species diversity for robust generative performance.

Overall Trends. The results confirm that TaxaDiffusion performs optimally with a balanced number of species, as seen in the low FID and high BioCLIP scores for *Actinopteri* and *Elasmobranchii*. However, as species diversity decreases, the model struggles to maintain comparable performance, suggesting that enhancing data augmentation or leveraging external sources could further benefit low-diversity classes.

In summary, the analysis highlights the strengths of TaxaDiffusion in generating high-fidelity, semantically aligned images for high-diversity classes and identifies areas for improvement, particularly for classes with limited species representation. Future work could explore advanced balancing techniques or domain adaptation strategies to address these challenges.

B.2. Results on the BIOSCAN-1M Insect Dataset

We further validate TaxaDiffusion using the BIOSCAN-1M Insect Dataset [3], which is a large-scale dataset designed to enable image-based taxonomic classification of insects. The dataset includes high-quality microscope images of insects, each annotated with taxonomic labels ranging from species to higher taxonomic ranks such as genus, family, order, and class. In addition to visual data, the dataset provides associated genetic information, such as DNA barcode sequences and Barcode Index Numbers (BINs), making it a valuable resource for biodiversity assessment and machine learning applications. However, the dataset poses challenges, such as a long-tailed distribution of classes and incomplete taxonomic labeling for many records.

The motivation for using this dataset stems from its po-

Table S2. Results evaluated at the class level for balanced subsets of the dataset. For classes with fewer than 400 species, all available species are included. We generate additional images per species to maintain balance. For classes with more than 400 species, a subset of 400 species was sampled. We report the FID for visual fidelity, LPIPS for perceptual similarity, and BioCLIP for semantic alignment.

Class name	# of Generated Images	# of Species	Images per Species	FID ↓	LPIPS ↓	BioCLIP ↑
Actinopteri	6000	400	15	27.75	0.7236	13.78
Elasmobranchii	6000	400	15	38.67	0.7324	12.78
Dipneusti	250	5	50	159.77	0.6910	8.87
Myxini	1050	21	50	119.23	0.7733	11.48
Cladistii	650	13	50	97.92	0.7226	8.34
Holocephali	1450	29	50	109.48	0.7364	13.10
Petromyzonti	1200	24	50	93.14	0.7954	9.91

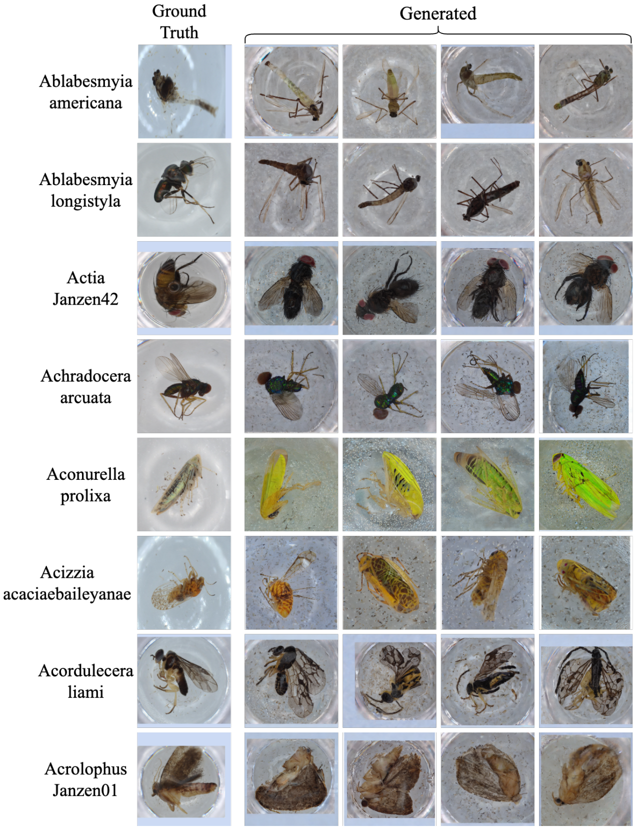
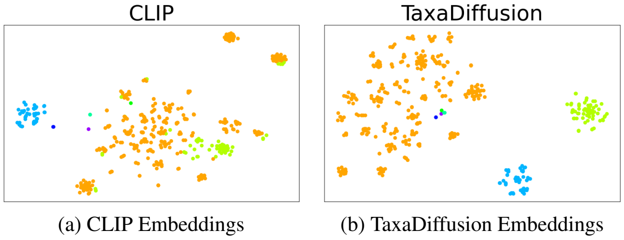


Figure S1. TaxaDiffusion results on BIOSCAN-1M dataset [3].

tential to validate the capability of our proposed TaxaDiffusion to handle highly diverse and fine-grained taxonomic classes. To ensure consistency with our training pipeline and focus on hierarchical taxonomic generation, we filtered the BIOSCAN-1M dataset to include only records with complete taxonomic information from species to class. After filtering, we retained 84,443 records, ensuring that each sample included annotations for species, genus, family, order, and class levels.

We train TaxaDiffusion on the filtered BIOSCAN-1M dataset for 50 epochs at each taxonomic level, following the same progressive training regime. The progressive training process leverages hierarchical taxonomic relationships, starting from higher levels (e.g., class and order), to capture shared traits before specializing at the species level. This approach allows the model to generate images that are not only visually accurate but also semantically aligned with the hierarchical structure of the taxonomy.



(a) CLIP Embeddings (b) TaxaDiffusion Embeddings

Figure S2. **CLIP vs. TaxaDiffusion embeddings.** We show t-SNE visualizations of CLIP and TaxaDiffusion embeddings. Different colors represent different `Class` level categories showcasing TaxaDiffusion learns embeddings that are more distinct and form well-separated clusters compared to CLIP embeddings.

The results, as shown in Figure S1, demonstrate the strength of our method in generating high-quality and taxonomically coherent insect images across various levels of the taxonomy. By effectively utilizing hierarchical information and addressing class imbalance through progressive training, TaxaDiffusion excels in fine-grained image generation for a diverse set of insect classes.

B.3. Comparison with State-of-the-Art Methods

Comparison with FineDiffusion [5] Figure S3 illustrates the results for a subset of species common to the iNaturalist and FishNet datasets. While FineDiffusion produces similar high-resolution images (512 x 512), the generated images fail to align with ground truth images, lacking species-specific details, whereas, TaxaDiffusion generates images that closely resemble ground truth, effectively capturing fine-grained characteristics unique to each species. This demonstrates the effectiveness of TaxaDiffusion, achieving

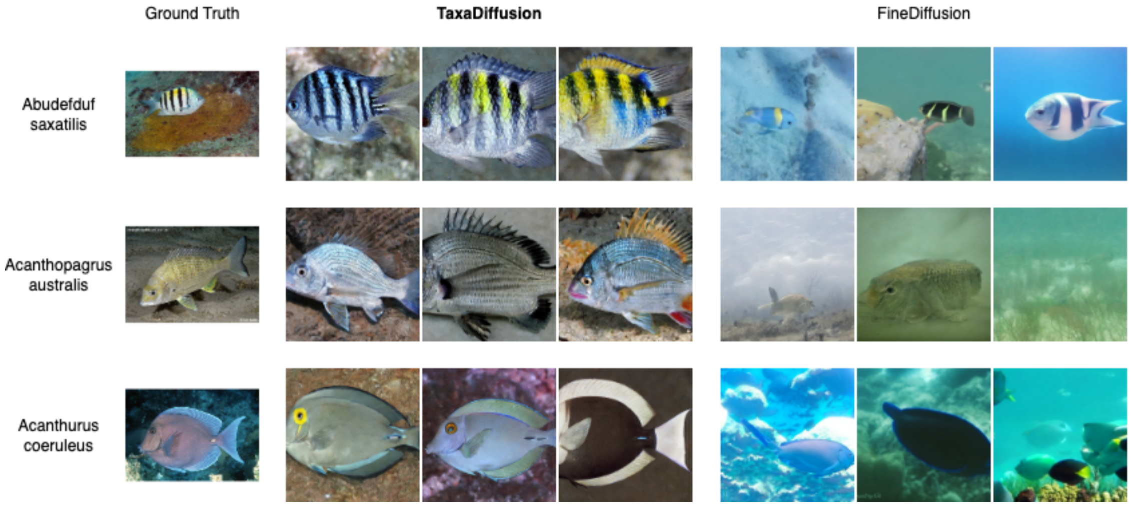


Figure S3. **Comparison with SOTA.** Qualitative comparison between TaxaDiffusion and FineDiffusion.

superior performance even when compared to FineDiffusion, which utilizes DiT-XL/2 model, a framework shown to outperform U-Net-based architectures [6].

B.4. Embedding Comparison

We investigate the quality of embeddings of TaxaDiffusion with the CLIP embeddings using t-SNE visualizations. Figure S2 shows the embeddings at the Family taxonomic levels for both models with diverse colors representing the unique Class level. TaxaDiffusion exhibits well-defined clusters, showing a clear separation between different Class groups, whereas CLIP embeddings used by baselines exhibit overlapping areas, particularly between the orange and green clusters. This hinders its ability to generate images that capture species-specific traits.

B.5. Further Analyses on Guidance Scale

To understand the effect of the guidance scale on image quality and trait specificity, we evaluate our model across a range of guidance scale values: $0, 2, 4, 6, 8, 10$. The guidance scale directly influences the intensity of the taxonomy-driven conditioning, modulating the balance between the generality and specificity of the generated samples. Figure S5 illustrates representative samples generated using different guidance scales, highlighting the progressive enhancement in fine-grained details and the potential trade-offs at extreme values.

Our experiments reveal a trend: at lower values (0 and 2), the generated images exhibit limited alignment with the fine-grained taxonomic traits. As the scale increases, the model's focus on species-specific details improves, with noticeable enhancements in morphological accuracy and distinctiveness at a guidance scale of 6. Beyond this point, higher values (8 and 10) result in overly restrictive guidance, which limits the diversity of the generated samples and sometimes introduces artifacts due to excessive specificity. Based on these findings, we selected a guidance scale of 6 for all subsequent experiments, achieving an optimal balance between fidelity and trait specificity.

B.6. Attention Visualizations

In Figure S6, we show attention maps of each level on two generated fishes *Acestrorhynchus falcatus* and *Acestrorhynchus falcirostris* with the same Genus but different Species. Our method captures shared traits at higher levels and distinct features at the Species level – the second case focuses more on the tail, less on the head.

C. Limitations

While our proposed TaxaDiffusion demonstrates strong generative capabilities across a wide range of taxonomic classes, certain limitations arise in scenarios where the dataset lacks sufficient diversity or representation. One notable case is the class *Dipneusti* within the FishNet dataset. This class contains only 5 species, with a total of 19 samples, which poses a significant challenge for our progressive diffusion framework.

The progressive training approach relies on transferring shared traits and information from higher taxonomic levels to refine the generative process at the species level. However, the extreme sparsity in the *Dipneusti* class limits this transfer, resulting in suboptimal generations for this group. As shown in Figure S4, the generated samples for *Dipneusti* often fail to replicate the fine-grained traits and diversity observed in the ground truth. Instead, the model struggles to generalize effectively due to the insufficient data representation, highlighting the importance of adequate class diversity

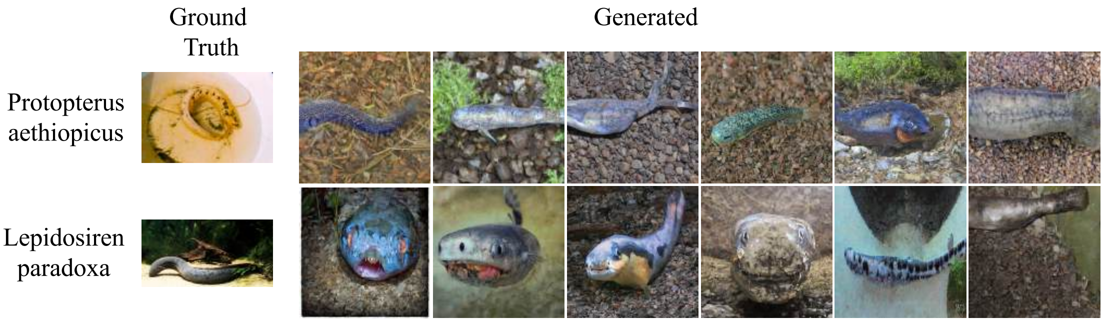


Figure S4. **Failure cases.** Examples of poor generations for the *Dipneusti* class due to insufficient data diversity and representation.

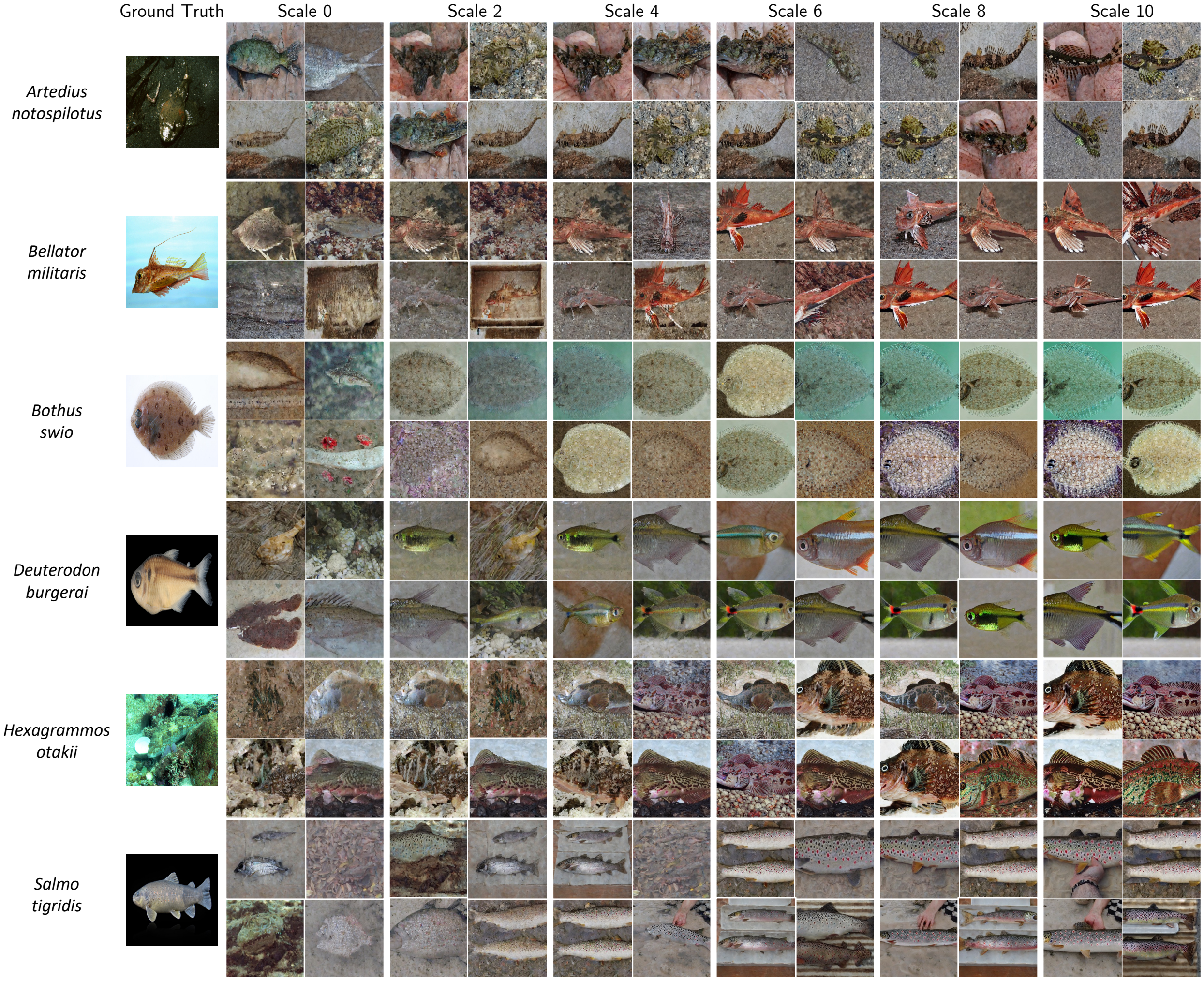


Figure S5. **Scale variations in TaxaDiffusion**. We conduct an analysis of different scale factors used for TaxaDiffusion. Images for each of the six species are generated with 250 inference steps with different TaxaDiffusion scale factors of 0, 2, 4, 6, 8, and 10 to show the balance between the generality and specificity of the generated samples compared to the ground truth training images.

for hierarchical approaches like ours.

Addressing such limitations would require strategies to augment the training process, such as incorporating additional data, leveraging synthetic data generation (using some methods like DreamBooth [8] to generate images for each specific species), or employing data augmentation

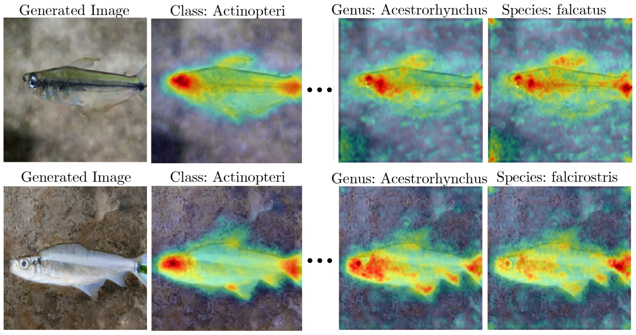


Figure S6. **Attention maps** of two species *Acestrorhynchus falcatus* and *Acestrorhynchus falcirostris* having the same `Genus`

techniques. Additionally, methods that enhance learning under extreme class imbalance or sparsity, such as few-shot learning or domain adaptation, could be explored to mitigate these challenges.

Trait discovery is a challenging problem, particularly when working with datasets like FishNet, which predominantly captures images of fish species in natural environments. Many fish species blend into their surroundings through camouflage or appear in diverse orientations, making it difficult to identify traits, especially for species with limited training images. To address these challenges, we plan to incorporate datasets with plain backgrounds, such as those commonly found in museum collections [1, 2], where controlled settings provide clearer views of species traits [4].

D. Ethics and Social Impacts

We focus on advancing generative models to progressively generate images of animal species, aiming to accelerate scientific discovery and research. Our work does not involve sensitive or human data, and we do not foresee any ethical concerns or negative societal impact associated with TaxaDiffusion or the results obtained.

References

[1] idigbio. *http://www.idigbio.org/portal*, 2020. 6

[2] Inhs collections data. *http://biocoll.inhs.illinois.edu/portal/index.php*, 2022. 6

[3] Zahra Gharaee, ZeMing Gong, Nicholas Pellegrino, Iuliia Zarubiieva, Joakim Bruslund Haurum, Scott Lowe, Jaclyn McKeown, Chris Ho, Joschka McLeod, Yi-Yun Wei, et al. A step towards worldwide biodiversity assessment: The bioscan-1m insect dataset. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[4] Kazi Sajeed Mehrab, M Maruf, Arka Daw, Abhilash Neog, Harish Babu Manogaran, Mridul Khurana, Zhenyang Feng, Bahadir Altintas, Yasin Bakis, Elizabeth G Campolongo, et al. Fish-vista: A multi-purpose dataset for understanding & identification of traits from images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24275–24285, 2025. 6

[5] Ziying Pan, Kun Wang, Gang Li, Feihong He, and Yongxuan Lai. Finediffusion: Scaling up diffusion models for fine-grained image generation with 10,000 classes. *arXiv preprint arXiv:2402.18331*, 2024. 3

[6] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 4

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[8] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 5