Not All Birds Look The Same: Identity-Preserving Generation For Birds

Supplementary Material

Dataset	# Pairs	# Classes	# Species
NABLA	4759	539	401
iNat-Seen	677	395	395
iNat-Unseen	396	396	396

Table A1. **Dataset statistics for NABLA, iNat-Seen, and iNat-Unseen.** Each pair contains two unique images, meaning the number of images is just twice the number of pairs.

Control	Backbone	Train	Test	DINO
Depth	Kontext	Short	Short	0.77
Depth	Kontext	Short	Long	0.75
Depth	Kontext	Long	Short	0.77
Depth	Kontext	Long	Long	**0.77**
Depth	Schnell	Short	Short	0.53
Depth	Schnell	Short	Long	0.53
Depth	Schnell	Long	Short	0.56
Depth	Schnell	Long	Long	**0.59**

Table A2. **Caption length ablation.** Training and evaluating on the long captions works better.

A. Additional Dataset Information

A.1. NABLA Annotation

A small group of birdwatcher volunteers with 8+ years of active field experience annotated the NABLA dataset. For each datapoint, a subject image was selected at random and then 6 images of the same class were displayed to the annotator (Figure A1). The annotator was given the task to select the image where the individual appears the most similar to the subject image. More specifically, we asked the annotators to evaluate the similarity of the birds in the two images using the following criteria:

1. Plumage: Do the color patterns and textures on the surfaces of the birds appear the same?
2. Structure: Do the body shapes and proportions (*e.g.* bill length, tail length, head shape, *etc.*) of the two birds appear the same?

We asked annotators to choose images which contained individuals with matching plumage and structure such that they could plausibly be the same bird in a different setting. Some examples of look-alike and non-look-alike pairs are given in Figure A9, along with reasons for why the negative pairs would not be considered a match. Note these criteria do not include matching bird pose or background lighting conditions, since we explicitly wanted a diversity of conditions for generation. If none of the candidate images were close enough, the annotator was given the option to shuffle the candidate images or to skip the subject image. This was repeated until approximately 5-10 images were selected for each class.

A.2. Dataset Statistics

In Table A1 we outline the basic statistics for the NABLA, iNat-Seen, and iNat-Unseen datasets. We see that NABLA contains 539 classes across 401 species, meaning over 100 species have different classes based on age, sex, or breeding status. Furthermore, a majority of classes in NABLA have at least 10 image pairs, shown in the left plot of Figure A2.

To create the iNaturalist datasets, we queried the iNaturalist API to get recent observations based on species name for research quality data. For iNat-Seen, we used the species list of NABirds and sampled 1-2 observations per species. For iNat-Unseen, we used the Aves species list from the 2017 iNaturalist competition, excluding the birds in NABirds, and sampled 1 observation per species. For observations with more than 2 images, we randomly selected 2 images to serve as the representative pair.

In the right plot of Figure A2 we see NABLA images generally have larger subjects than iNaturalist data. We consider this to be correlated with image quality, indicating NABLA images to generally be higher quality than our iNat datasets.

A.3. Caption Generation

We explore two different captioning modes for training and evaluation. In both cases we use Qwen-2.5 VL for generating captions on each image, but vary the prompt to get short captions and long captions. The prompt for each setting is given in Table A3 along with examples for each one. We found in Table A2 the long captions worked better but the differences were quite small overall.

A.4. Additional Visualizations

We show additional image pairs from NABLA, iNat-Seen, and iNat-Unseen in Figures A6, A7, and A8, respectively. In Figure A9, we highlight potential mismatched pairs and compare them to corresponding pairs in NABLA. Individual differences in plumage can occur even within a class which leads to inaccurate evaluation on the identity-preserving task.

A.5. Other Dataset Comparisons

We also briefly discuss differences between NABLA and other existing datasets which were not highlighted in the

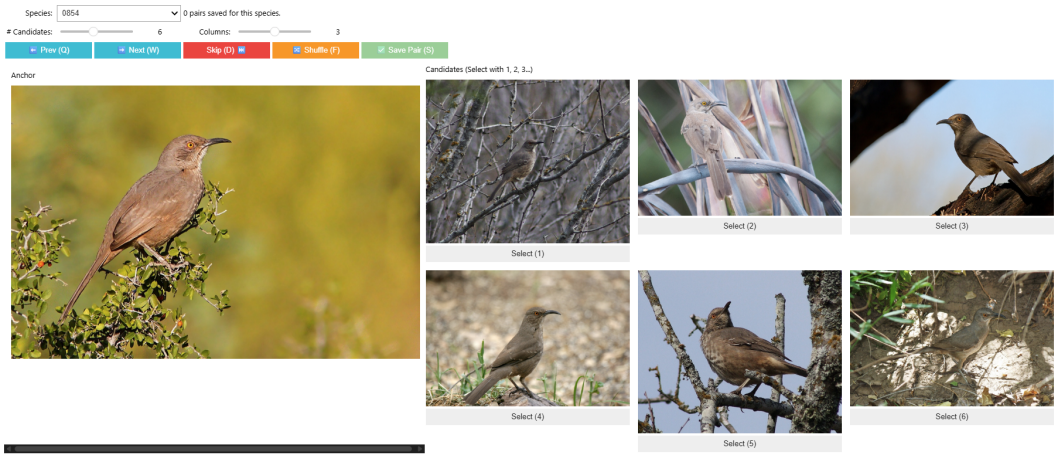

Figure A1. **NABLA annotation widget.** Annotators are given the task of selecting an image where the individual present in the image looks the same as the individual in the anchor image. By default, 6 candidate images are shown at random from the same class. Annotators have the option to shuffle the candidates or skip the anchor image.

	Short Caption	Long Caption
Prompt	Look at this image and describe where the SUBJECT_ITEM : 'bird' is placed. Be *very* brief but do not miss elements EXCEPT the SUBJECT_ITEM. DO NOT DESCRIBE OR MENTION THE SUBJECT_ITEM. You should output starting with: "Place it in" or + "Place it on."	Look at this image and describe where the bird is and what the background of the image is. DO NOT DESCRIBE OR MENTION THE POSE OR APPEARANCE OF THE BIRD.
	"Place it on a rocky surface near water."	"The bird is on a rocky surface near a body of water, which appears to be calm and overcast. The background consists of the water and the sky, which is gray and cloudy."
	"Place it on leaves."	"The bird is standing on the ground, surrounded by a bed of dried leaves. The background consists of scattered leaves in various shades of brown and green, creating a natural forest floor setting."

Table A3. **Sample captions from NABLA generated using Qwen2.5 VL.** Each image has two captions, one long and one short. We tried training and evaluating the depth and keypoint models using short and long captions but generally found better results using the long caption.

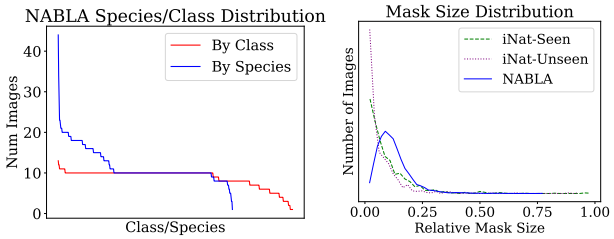

Figure A2. **NABLA, iNat-Seen, and iNat-Unseen statistics.** Left: Most classes in NABLA have 10 pairs. Right: Images in NABLA typically have larger subjects than iNat-Seen or iNat-Unseen images.

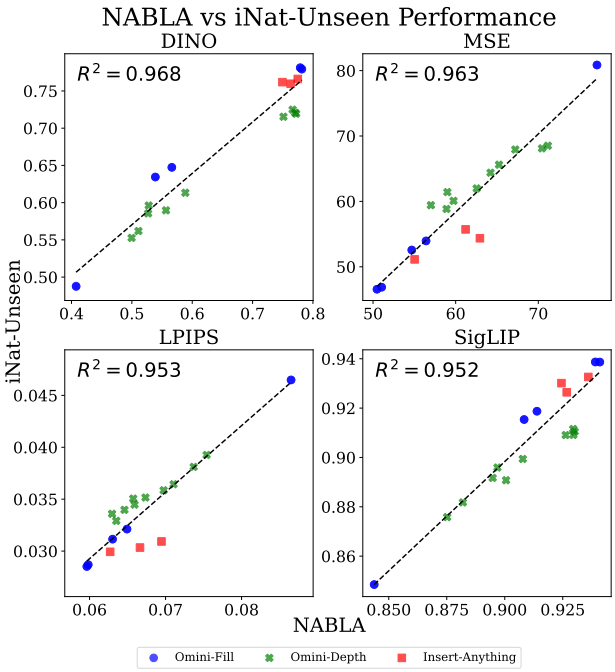

Figure A3. **NABLA and iNat-Unseen Performance Correlation.** Comparison of average model performance on NABLA and iNat-Unseen data across all trials.

main text. DreamBench++ [33] is a human-aligned benchmark but focuses on the quality of creative synthetic generations to be assessed using GPT-4o as opposed to realism and accuracy. DeepFashion2 [11], PODS [47], and CUTE [19] are all true same-subject datasets but focus on everyday objects and clothes which have relatively simple pose variation. WildlifeReID [1] and PetFace [42] are both re-identification datasets on animals. WildlifeReID compiles various re-identification datasets into a single benchmark but this only spans 33 species across the entire animal kingdom. PetFace impressively spans over 250 thousand individuals across 13 families, but these photos are facially-focused for the purposes of re-identification and focuses on commonly-owned pets such as cats and dogs. Compared

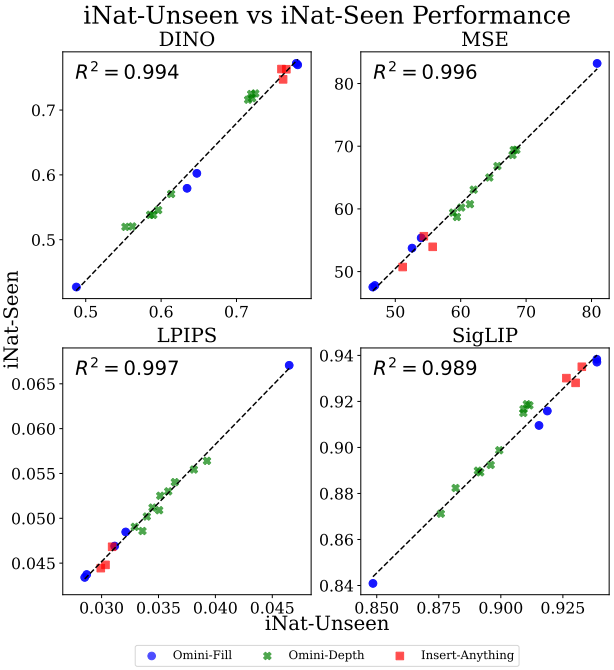

Figure A4. **iNat-Unseen and iNat-Seen Performance Correlation.** Comparison of average model performance on iNat-Unseen and iNat-Seen data across all trials.

Figure A5. **Mixed depth + fill control.** The mixed control mode is a simple combination of the depth map and background fill image, where the mask is replaced with the pixels from the depth map.

to iNaturalist, we found that these two re-identification datasets were limited in terms of image resolution, pose and lighting variation, and species diversity, especially for evaluating image generation tasks.

B. Training Hyperparameters

B.1. Basic Hyperparameters

For both OminiControl [48] and Insert Anything [43] we used the Prodigy optimizer [23] with a learning rate of 1, weight decay of 0.01, and safeguard warmup and bias correction set to true. On OminiControl experiments we ran with a batch size of 24 while for Insert Anything we used a batch size of 20 with gradient checkpointing for both experiments. We match the LoRA settings of OminiControl and Insert Anything training procedures exactly. At evaluation time we use guidance scale of 2.5 for OminiControl.

B.2. Keypoint Control

We define a sparse skeleton on the NABirds keypoints. The skeleton is defined with the following edges: (bill, crown), (crown, nape), (left eye, bill), (right eye, bill), (belly, breast), (breast, bill), (back, nape), (tail, back), (left wing, back), and (right wing, back). We explored a variety of joint and edge sizes. We found the best results were for a joint diameter of 15 pixels and edge width of 10 pixels.

B.3. Proprietary Model Settings

Nano Banana and GPT-4V were run on their public websites from Nov. 1-7'25 using the "Create Image" feature on Pro and default modes, respectively. The prompt and images are described in Figure 2 in the paper. The propriety models were provided the subject image, masked background, and the prompt "Please inpaint this bird into the pose given by the black mask."

C. Additional Results

C.1. Correlation Graphs

Performance correlation graphs have been also been generated for the NABLA and iNat-Unseen pair and iNat-Seen and iNat-Unseen pair in Figures A3 and A4, respectively. We see the correlation between the three datasets is strong, but the iNat-Seen and iNat-Unseen datasets have the most similar performance.

C.2. Generation Results

We show additional generation results on our best models for each control from NABLA, iNat-Seen, and iNat-Unseen in Figures A10, A11, and A12, respectively.

C.3. Individual-Birds Dataset Evaluation

In Table A4, we present results from evaluating our fine-tuned models on Individual-Birds [9] from val+test sets in WildlifeReID-10k [1]. For each individual, 3 random image pairs were selected for evaluation. Since many Zebra Finch images include multiple subjects, we excluded these from our evaluation. We observe very similar trends to iNaturalist — models trained on NABirds using proxy pairs show improved performance over baselines.

C.4. Mixed Fill + Depth Control

In Table A5, we present results after fine-tuning OminiControl + Flux-Kontext using a mixed depth + fill control mode. The control image is simply the fill image with the pixels from the depth map pasted into the bird mask, as in Figure A5. We observe very similar results between the fill, depth, and mixed models, indicating a single image combination of the two controls is insufficient for improvement.

Control	Arch	DINO ↑	SigLIP ↑	LPIPS ↓	MSE ↓
Fill	Ins-A*	0.77	0.93	0.021	21.6
Fill	Ins-A	0.81	0.95	0.018	20.0
Fill	Om-S*	0.28	0.79	0.050	22.3
Fill	Om-S	0.51	0.90	0.031	24.3
Depth	Om-S*	0.42	0.84	0.042	25.1
Depth	Om-S	0.47	0.87	0.033	26.6

Table A4. Evaluating our models on random pairs in Individual-Birds from val+test sets in WildlifeReID-10k. * indicates baseline. 3 random image pairs were selected per individual. Zebra Finches were excluded since many images captured multiple subjects.

Control	DINO ↑	SigLIP ↑	LPIPS ↓	MSE ↓
Depth	0.77	0.93	0.063	57.0
Fill	0.78	0.94	0.060	51.0
Depth + Fill	0.77	0.94	0.061	57.0

Table A5. Comparing the mixed control mode of depth + fill to existing results on Om-K. The results are similar to depth/fill only.

Figure A6. **Additional NABLA examples.** Additional test examples from NABLA, sampled randomly. Images center-cropped to square.

Figure A7. **Additional iNat-Seen examples.** Additional test examples from iNat-Seen, sampled randomly. Images center-cropped to square.

Figure A8. **Additional iNat-Unseen examples.** Additional test examples from iNat-Unseen, sampled randomly. Images center-cropped to square.

Figure A9. **Unfiltered NABirds vs NABLA.** Comparing unfiltered sample pairs from the NABirds test set to NABLA pairs of the same species. Unlike NABLA, pairs in NABirds can have significantly different plumages between the two images despite sharing the same class. This can lead to inaccurate evaluations for the identity-preserving generation task.

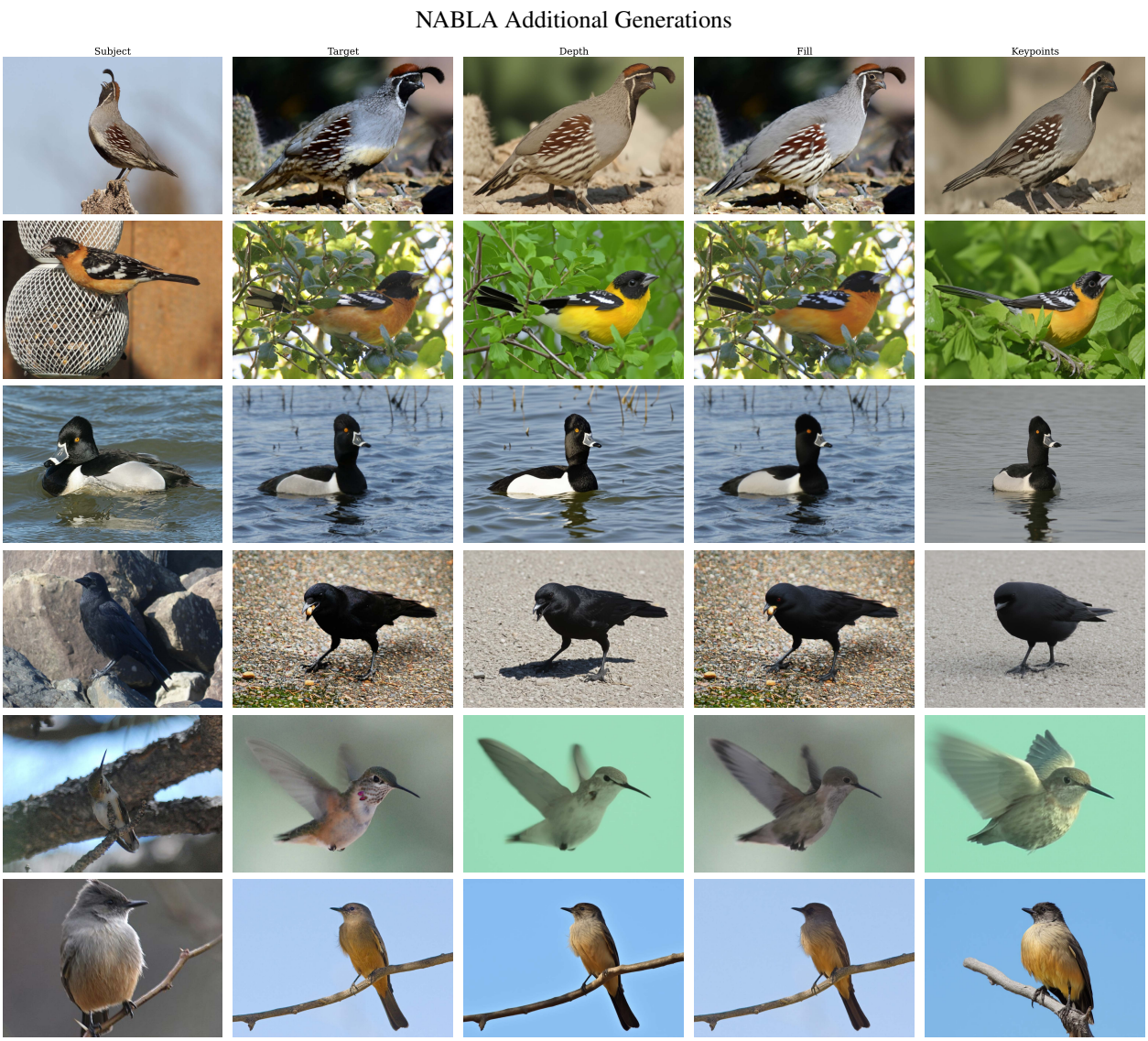

Figure A10. **Additional NABLA generations.** Additional test examples generations from NABLA, sampled randomly. Generations from best model of each control type. Images center-cropped to square.

Figure A11. **Additional iNat-Seen generations.** Additional test examples generations from iNat-Seen, sampled randomly. Generations from best model of each control type. Images center-cropped to square.

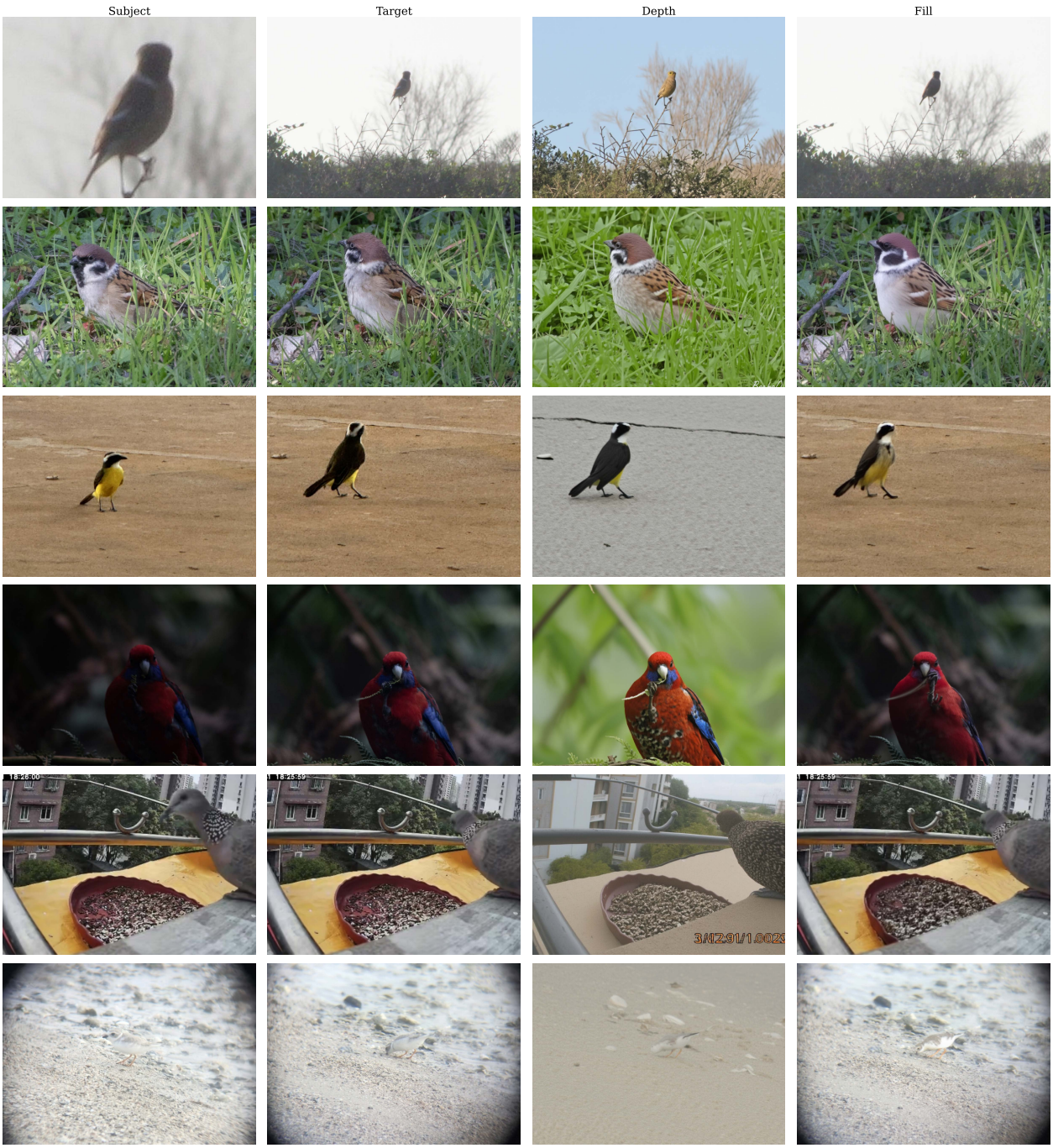

Figure A12. **Additional iNat-Unseen generations.** Additional test examples generations from iNat-Unseen, sampled randomly. Generations from best model of each control type. Images center-cropped to square.