Supplementary: Discriminative Class Tokens for Text-to-Image Diffusion Models

Idan Schwartz[1]* Vésteinn Snæbjarnarson[2]* Hila Chefer[1]
Serge Belongie[2] Lior Wolf[1] Sagie Benaim[2]
[1]Tel Aviv University [2]University of Copenhagen

In this supplementary, we extend the concepts presented in the main paper and describe the findings of our empirical analysis. The supplementary comprises the following subsections:

1. Sec. 1, batch size comparison
2. Sec. 2, examples of image generations with tokens trained different number of steps
3. Sec. 3, more details on the training setup
4. Sec. 4, capabilities of SD on CUB with failure cases
5. Sec. 5, FID comparison with SD
6. Sec. 6, qualitative samples

*Equal contribution.

1. Batch size ablation study

We experimented with the following batch sizes: $1, 2, 3, 4, 5, 6$. To measure the accuracy of each experiment, we generated 20 images with the trained discriminative token. We used three datasets: (1) CUB, (2) iNat21, and (3) ImageNet. Our study included a specific set of difficult classes, including those with ambiguous class names or those distinguished with fine details. The labels used for CUB are: *Chuck will Widow, Grasshopper Sparrow, Pied Kingfisher, Scissor tailed Flycatcher, Worm eating Warbler, Rhinoceros Auklet, Gadwall, Mourning Warbler, Spotted Catbird, Olive sided Flycatcher, Blue Grosbeak, Black billed Cuckoo, Brown Pelican, White necked Raven, Tree Swallow, Black throated Sparrow, Pied billed Grebe, Horned Grebe, Spotted Catbird, Heermann Gull*. The labels used for iNat are randomly selected: *grosbeak weaver, skeleton weed, ribbon jumping spider, common garden orb-web spinner, glass shrimp, crevice swift, eastern spiny lizard, double-barred finch, speckled swimming crab, new zealand cockle, common wall lizard, red-bordered pixie, old mans beard, rough-mantled dorid, green shore crab, ear fungus, european common cuttlefish, argentine black-and-white tegu, cassins vireo, steenbok*. The labels used for ImageNet are: *trilobite, kite, eft, bullfrog, Gila monster, crane, Japanese spaniel, drake, quail, Scotch terrier, curly-coated retriever, English setter, American chameleon, brass, bonnet, amphibian, ashcan, horizontal bar, iron, slot, spindle, sundial, steel drum, gong, coffee mug, vault, axolotl, tiger cat, bearskin*.

The results are summarized in Tab. 1. Using batch sizes 4-6 gives the best result depending on the dataset.

Batch Size	CUB Acc	iNat Acc	iNet Acc
Baseline	14.2	4.8	15.7
1	28.2	26.8	22.5
2	40.1	35.8	27.0
3	34.0	34.0	21.8
4	43.7	**29.6**	29.7
5	**54.2**	28.3	35.3
6	50.3	29.4	**39.3**

Table 1: Summary of accuracy and average steps needed over different batch sizes. SD accuracy is provided as a baseline.

In Fig. 1, we show comparisons over the CUB200 dataset for different batch sizes. In Fig. 2, the same sort of comparison is shown for iNat21. The images show that classifier guidance can adjust the appearance and add features to make the generated image more like a true sample. Note that the true images were hand-picked to be in a similar position as that in the generated images.

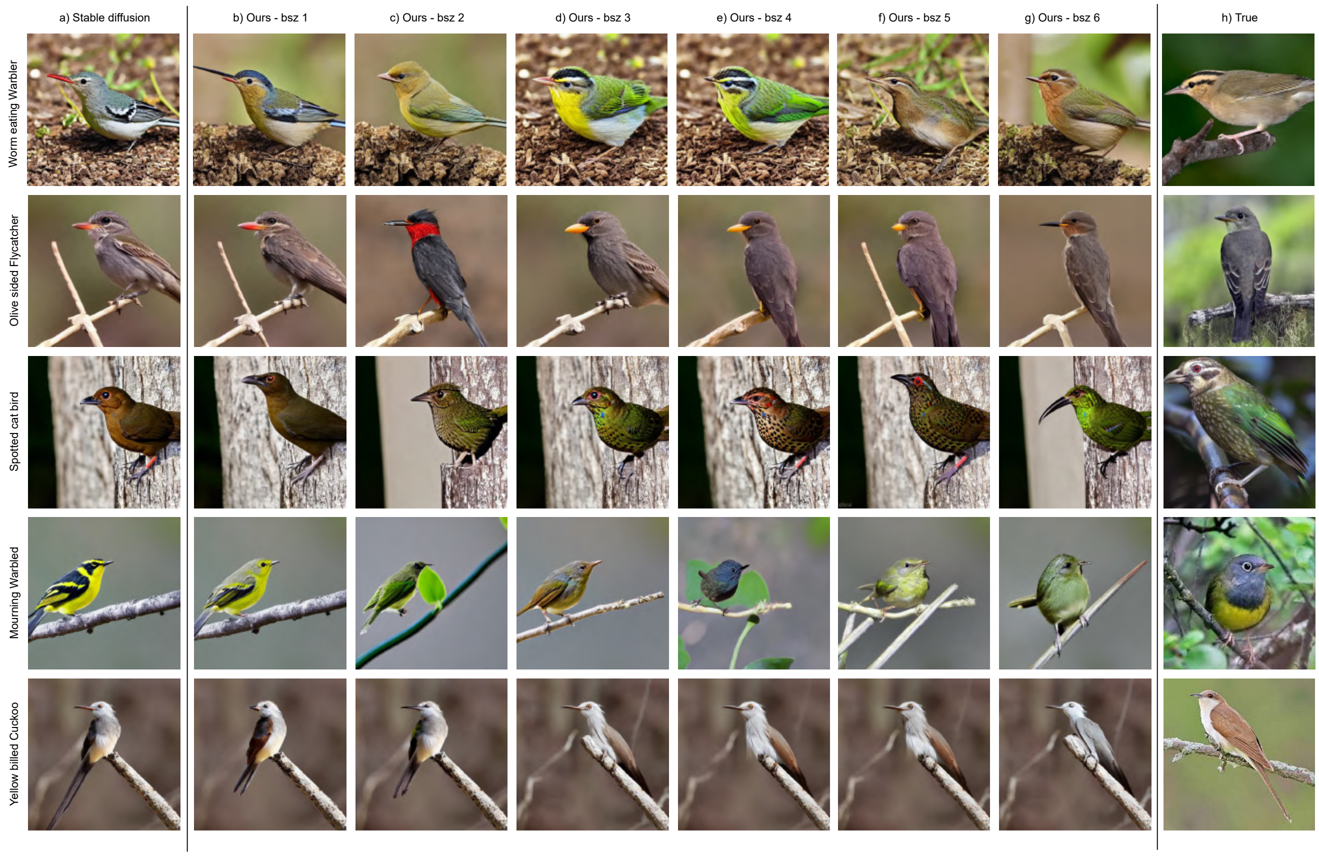


Figure 1: A comparison between Stable Diffusion, and six different batch sizes using CUB200 classifier guidance, as well as a real image.

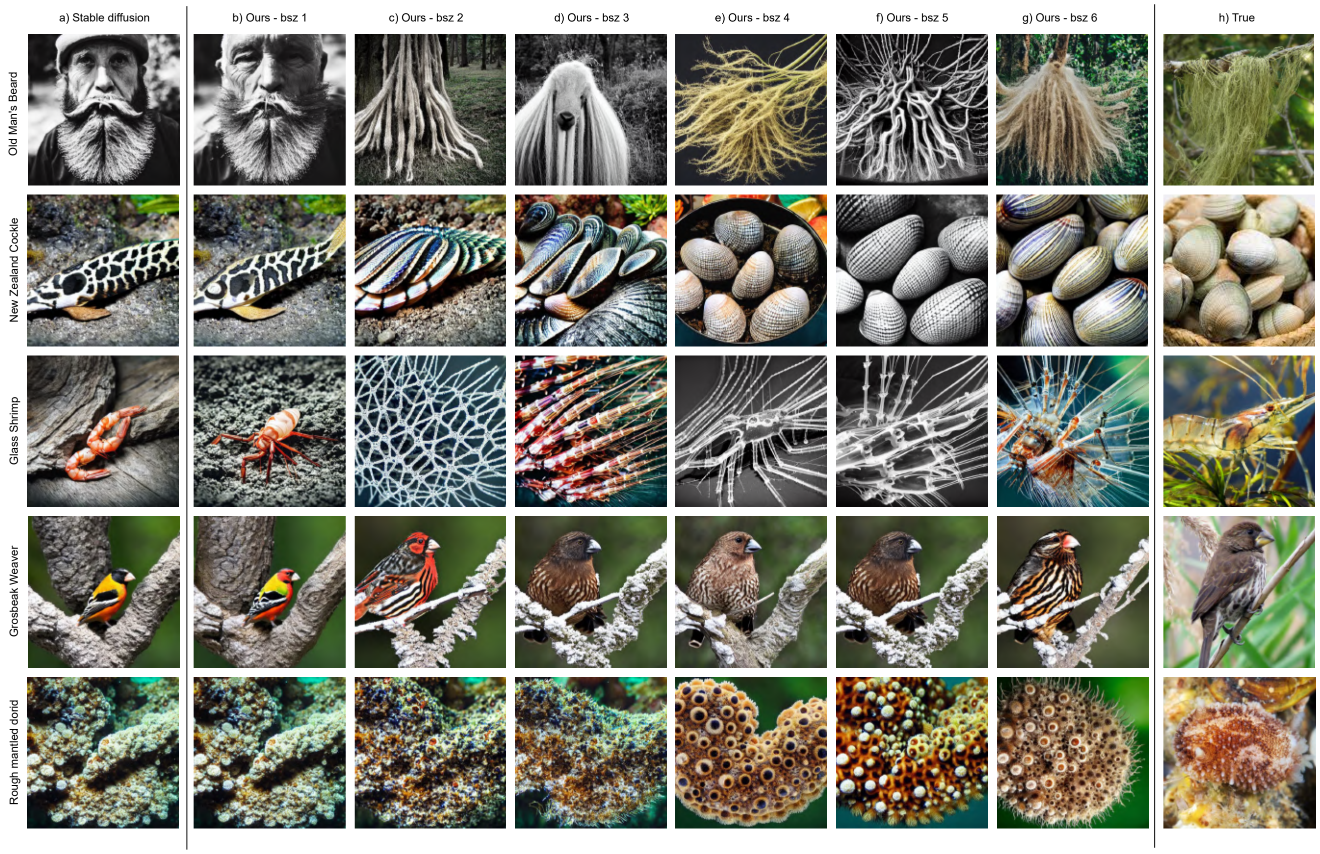


Figure 2: A comparison between Stable Diffusion, and six different batch sizes using iNat21 classifier guidance, as well as a real image.

2. Tokens from different training steps

Our proposed technique performs iterative updates on the class tokens, and the progress can be demonstrated by comparing images at each stage. For instance, we modified an image by utilizing class tokens from various training stages, as shown in Fig. 3, to depict this process. Our results indicate that at times the modified images become increasingly similar to authentic images in terms of structure, with less noticeable changes in color patterns. This finding suggests that shape may be more critical to classifiers, given the high variability in color patterns among objects and animals of the same class.

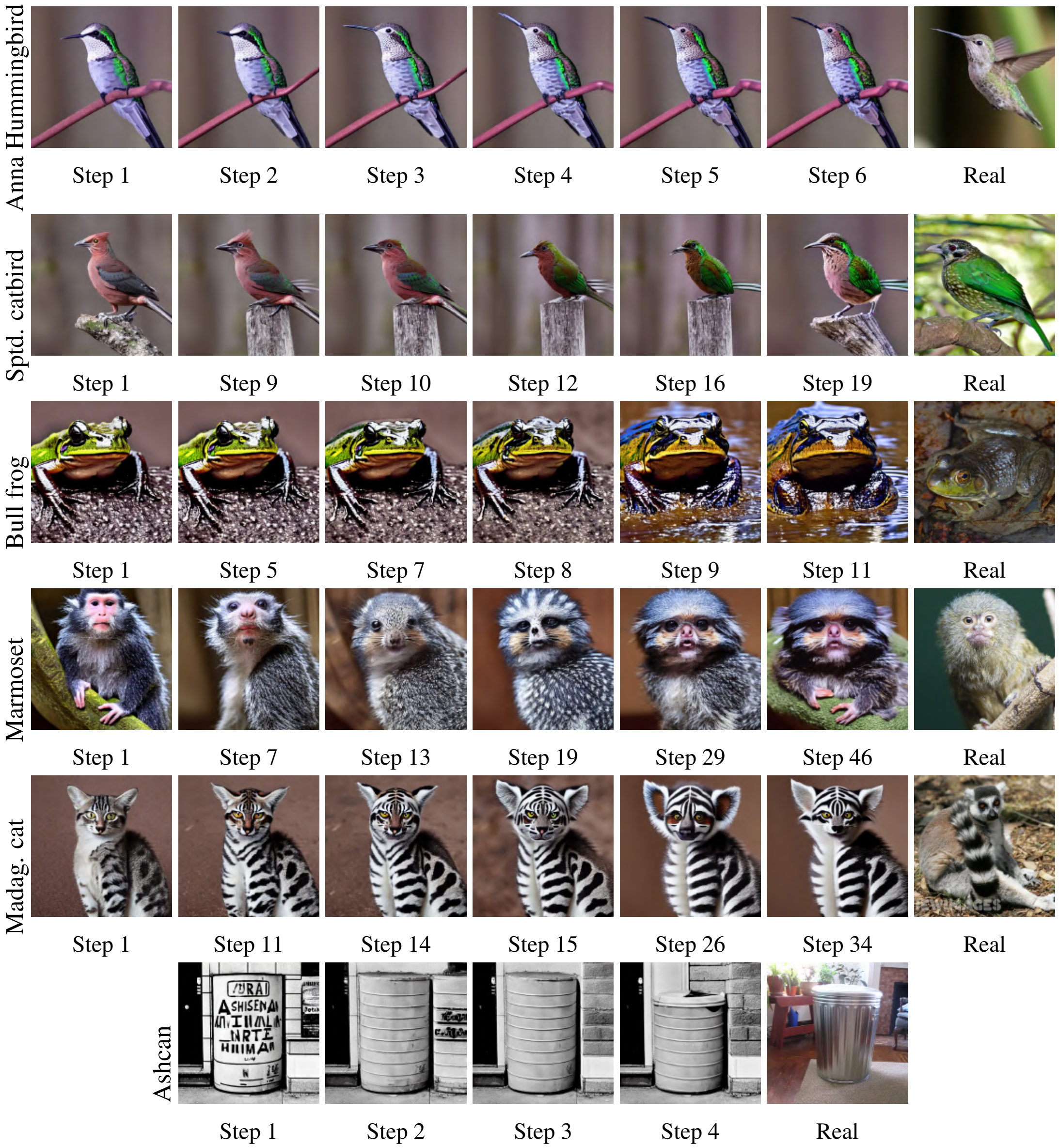


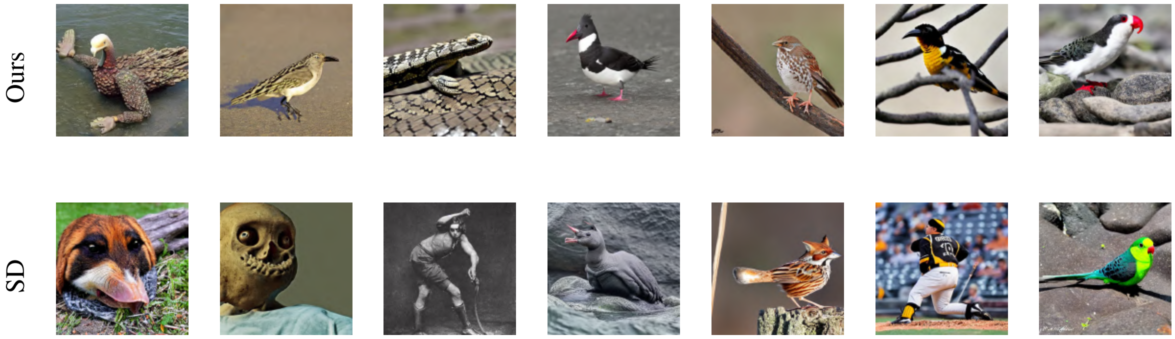
Figure 3: Intermediary results from class tokens fine-tuned for differing numbers of steps. Steps are chosen to highlight the overall changes.

3. More details on the training setup

In this section, additional information regarding our setup is provided. Our approach involved employing classifier-free guidance, with a scale of 7 and executing $T = 50$ denoising steps for the Stable Diffusion process. The generated images were created with a height and width of 512 pixels. Along with these parameters, there were other hyperparameters: (i) The learning rate; we used 0.0005. We recommend adapting it based on the batch size with the formula $0.00025 \cdot bsz$, where bsz indicates the batch size. (ii) Early stopping; we had three rules to stop: 1) If all the samples in the batch were classifier correctly. 2) If, for 20 steps, the loss has not improved, and at least 50% were classified correctly. 3) If we had trained for 200 steps. On average, seven steps are required, while the third rule appears in less than 1% of the examples. (iii) We used gradient clipping with a norm of one.

4. SD knowledge and failure cases in CUB

To evaluate the efficacy of SD in generating fine-grained categories, we conducted an analysis of SD's capabilities using the labels from the CUB dataset. This involved a thorough examination of all 20k images generated for CUB. Our investigation revealed that SD faced difficulties in generating images for seven out of the two hundred labels in CUB, as highlighted in Figure 4. Specifically, two species (Whip poor Will and Geococcyx) failed completely, with not a single image depicting a bird. 'Whip poor Will' generated images featuring whips and people, while 'Chuck Will Widow' captions resulted in 31 failing images that related mainly to hunting or dogs. Additionally, the Parakeet Auklet input rendered many images of its more colorful relatives, the Parakeets. Some species that mention other animals in their names resulted in mixed-up images that were still birdlike, such as the Rhinoceros Auklet (where all cases almost contained rhinoceros attributes) and the Fox sparrow (where 22 images contained fox-like attributes). Furthermore, three of the generated images for Scott Oriole contained scenes from baseball matches (the Baltimore Orioles is a baseball team). Our approach was successful in improving image generation for all classes except one: the 'Whip Poor Will' images ended up being of snakes. It appears that the specific camouflage of the bird in question misled both SD and our method.



(a) Chuck Will Widow (b) Geococcyx (c) Whip poor Will (d) Rhinoceros Auklet (e) Fox sparrow (f) Scott Oriole (g) Parakeet Auklet

Figure 4: Examples of SD failures for the CUB classes where the output does not represent the class as found in CUB, with our improvements on top. While 193 out of 200 bird names resulted in images of birds for SD, 6 classes were particularly troublesome and one happened to have three baseball-related pictures among the generations.

5. FID results comparison to SD

To further confirm the effectiveness of our method compared to SD, we employed the Frechet Inception Distance (FID) [1] to evaluate the similarity of the generated images to the real images in the CUB, iNat21, and ImageNet datasets. FID is a metric that assesses the quality of generated images by comparing them to a baseline dataset through the analysis of activations in classifiers. A lower FID score indicates a higher degree of resemblance between the generated and real images. We computed FID scores using two classifier models: one trained on ImageNet and another trained on CUB200. The FID results are presented in Table 2. Our findings suggest that our approach produces images that more accurately resemble the real images compared to those generated using only SD, as demonstrated by the lower FID scores for the InceptionV3 and CUB activations.

Dataset	Method	IncV3	CUB / iNat
CUB200	SD	14.7	116.8
CUB200	SD + Guidance	13.3	90.6
iNat	SD	83.7	225.6
iNat	SD + Guidance	82.4	207.3
ImageNet	SD	23.0	–
ImageNet	SD + Guidance	22.4	–

Table 2: FID scores, for all generated datasets compared to the real dataset, where features used for calculating the FID statistics are extracted from an Inception-v3 classifier trained on ImageNet, and classifiers trained on CUB200 / iNat. Lower is better.

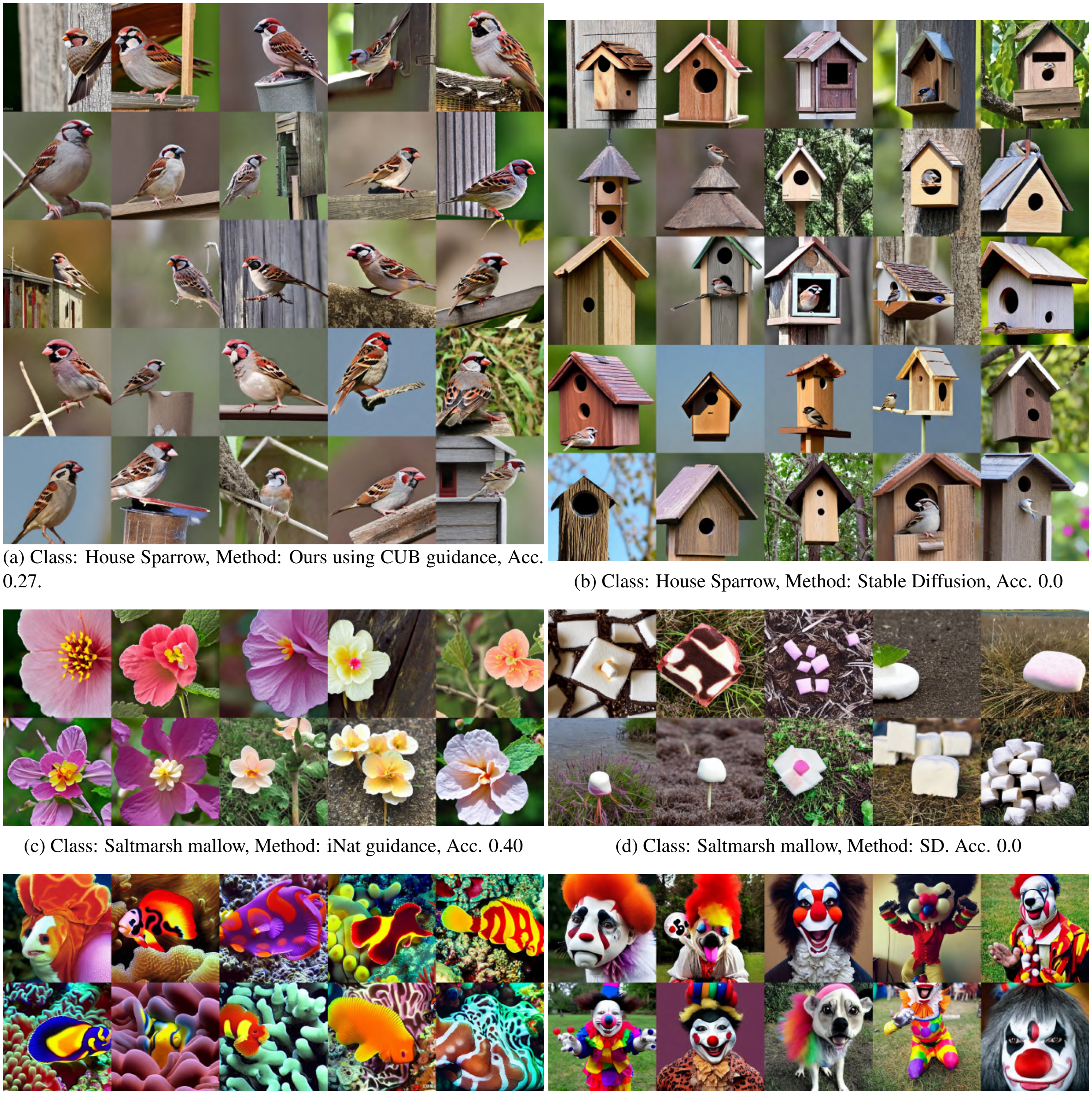
6. Qualitative study

In this section, we provide more qualitative results.

We show examples of ambiguous classes in Fig. 5 and Fig. 6. We note that even after our method's improvements, some may retain some characteristics of the original ambiguous image. House sparrows, for example, are often depicted in birdhouses, and Crane bird images still might depict a crane.

In Fig. 9, we show examples of coarse-grained classes from ImageNet. Images include various animals, everyday life objects, and places. With the imagenet classifier, we can improve fine details from dog faces to the style of a coffee mug. Our next step is to assess fine detail abilities. We show in Fig. 7 accuracy improvements using CUB and iNat datasets. Notably, the classifier pays careful attention to details. For instance, in the Red Sea Urchin, the red bubble in the center is incorrect, and in the Yellow-pine Chipmunk, the lines on his back are different.

Figure 5: Examples of ambiguous class names from CUB and iNat.



(a) Class: House Sparrow, Method: Ours using CUB guidance, Acc. 0.27.

(b) Class: House Sparrow, Method: Stable Diffusion, Acc. 0.0

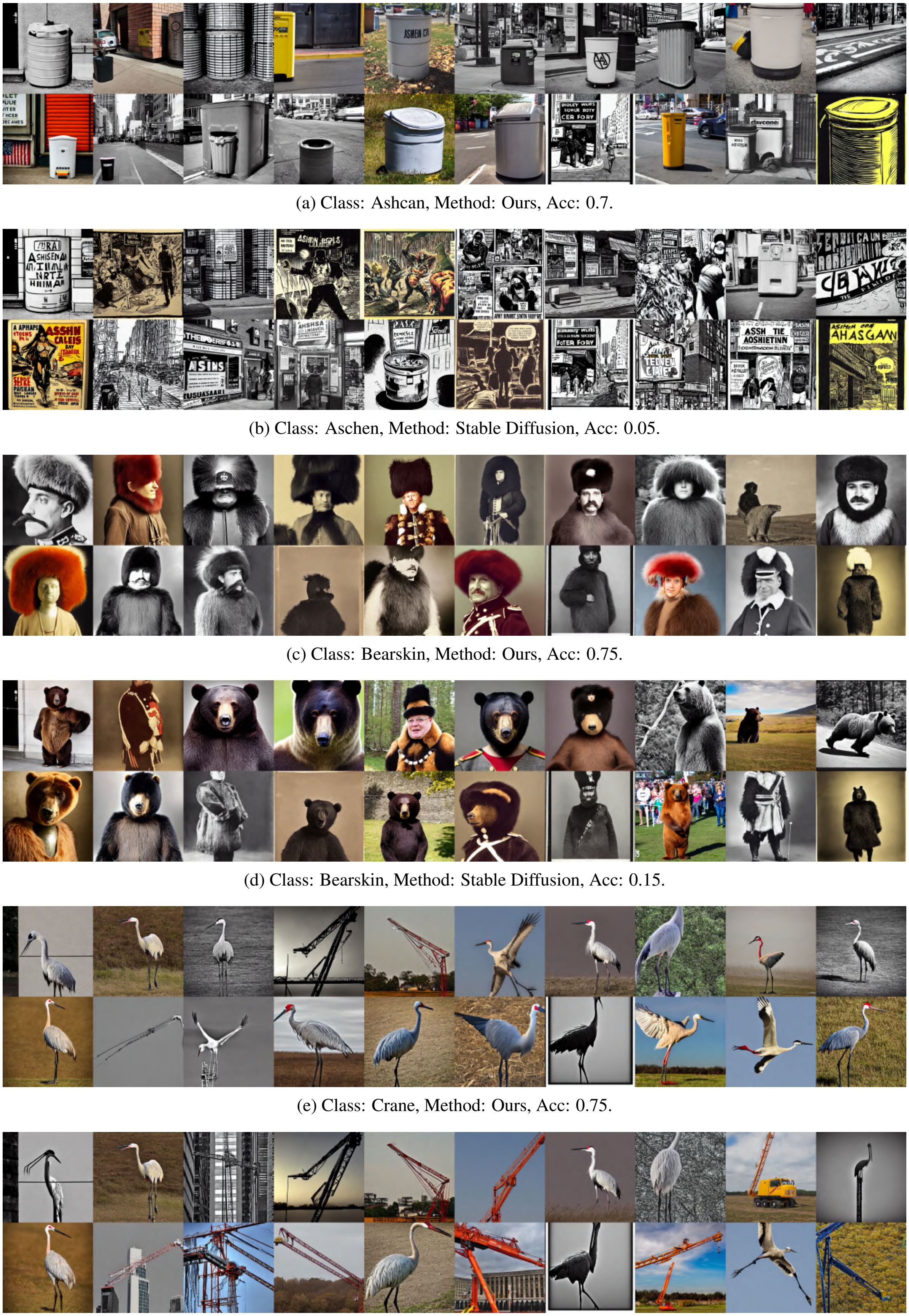
(c) Class: Saltmarsh mallow, Method: iNat guidance, Acc. 0.40

(d) Class: Saltmarsh mallow, Method: SD. Acc. 0.0

(e) Class: Clown doris, Method: Ours using iNat guidance, Acc 0.0

(f) Class: Clown doris, Method: Stable Diffusion. Acc 0.0

Figure 6: Examples of ambiguous class names from ImageNet.



(a) Class: Ashcan, Method: Ours, Acc: 0.7.

(b) Class: Aschen, Method: Stable Diffusion, Acc: 0.05.

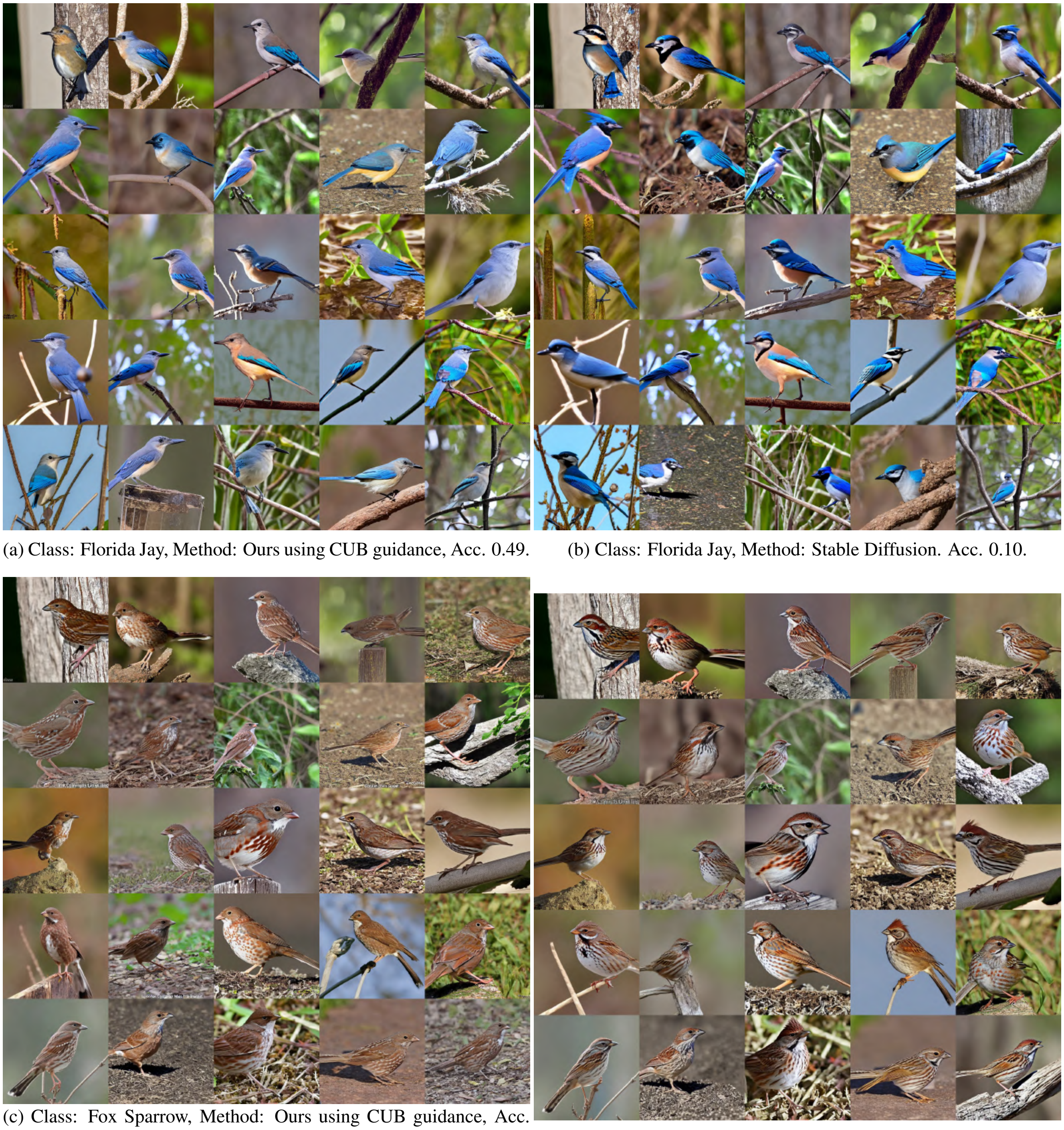
(c) Class: Bearskin, Method: Ours, Acc: 0.75.

(d) Class: Bearskin, Method: Stable Diffusion, Acc: 0.15.

(e) Class: Crane, Method: Ours, Acc: 0.75.

(f) Class: Crane, Method: Stable Diffusion, Acc: 0.25.

Figure 7: Images generated based on CUB classes. We show results with our discriminative token and vanilla Stable Diffusion. We provide the accuracy of the classification for each batch.

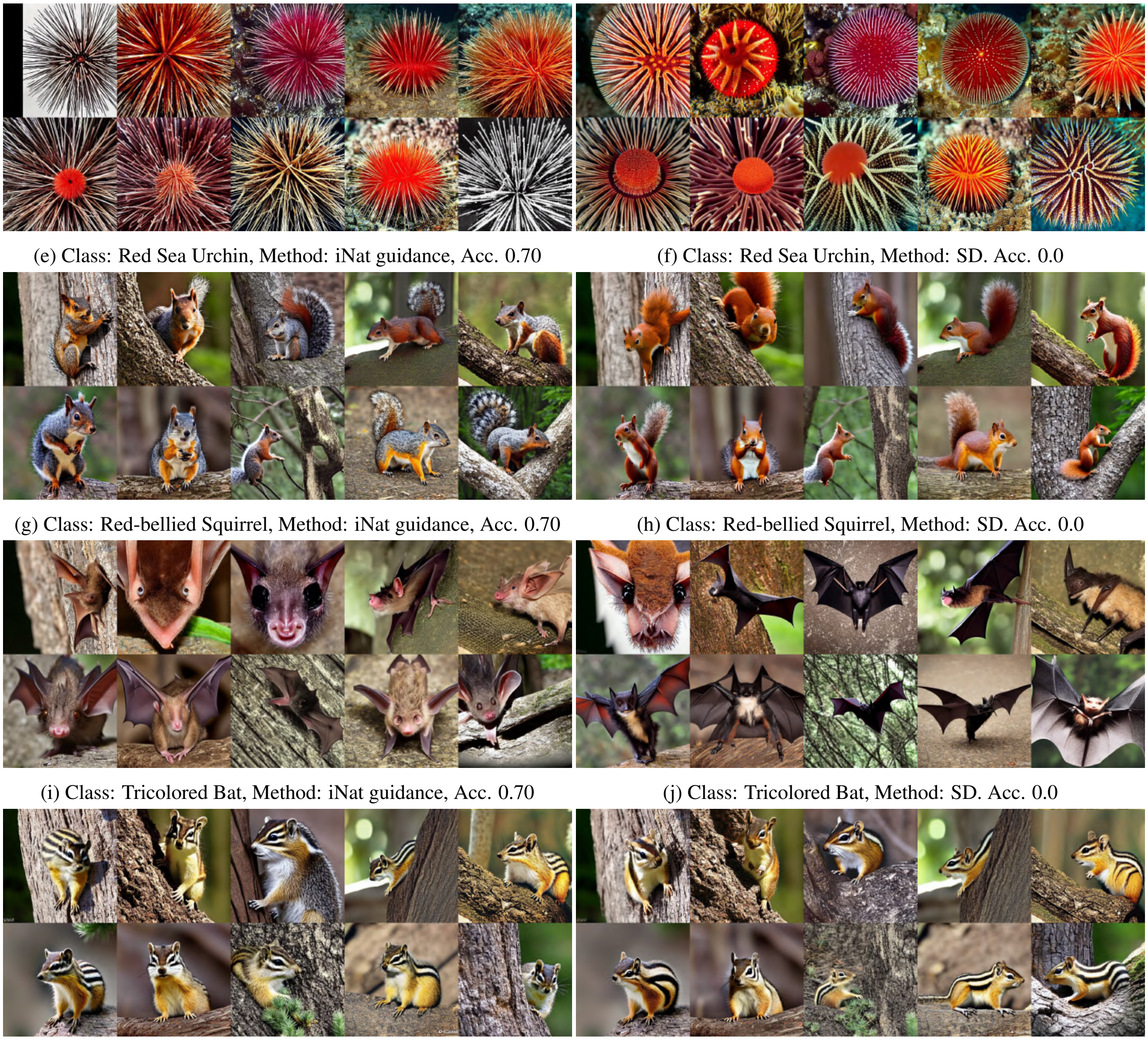


(a) Class: Florida Jay, Method: Ours using CUB guidance, Acc. 0.49.

(b) Class: Florida Jay, Method: Stable Diffusion. Acc. 0.10.

(c) Class: Fox Sparrow, Method: Ours using CUB guidance, Acc. 0.72.

(d) Class: Fox Sparrow, Method: Stable Diffusion. Acc. 0.9.



(e) Class: Red Sea Urchin, Method: iNat guidance, Acc. 0.70

(f) Class: Red Sea Urchin, Method: SD. Acc. 0.0

(g) Class: Red-bellied Squirrel, Method: iNat guidance, Acc. 0.70

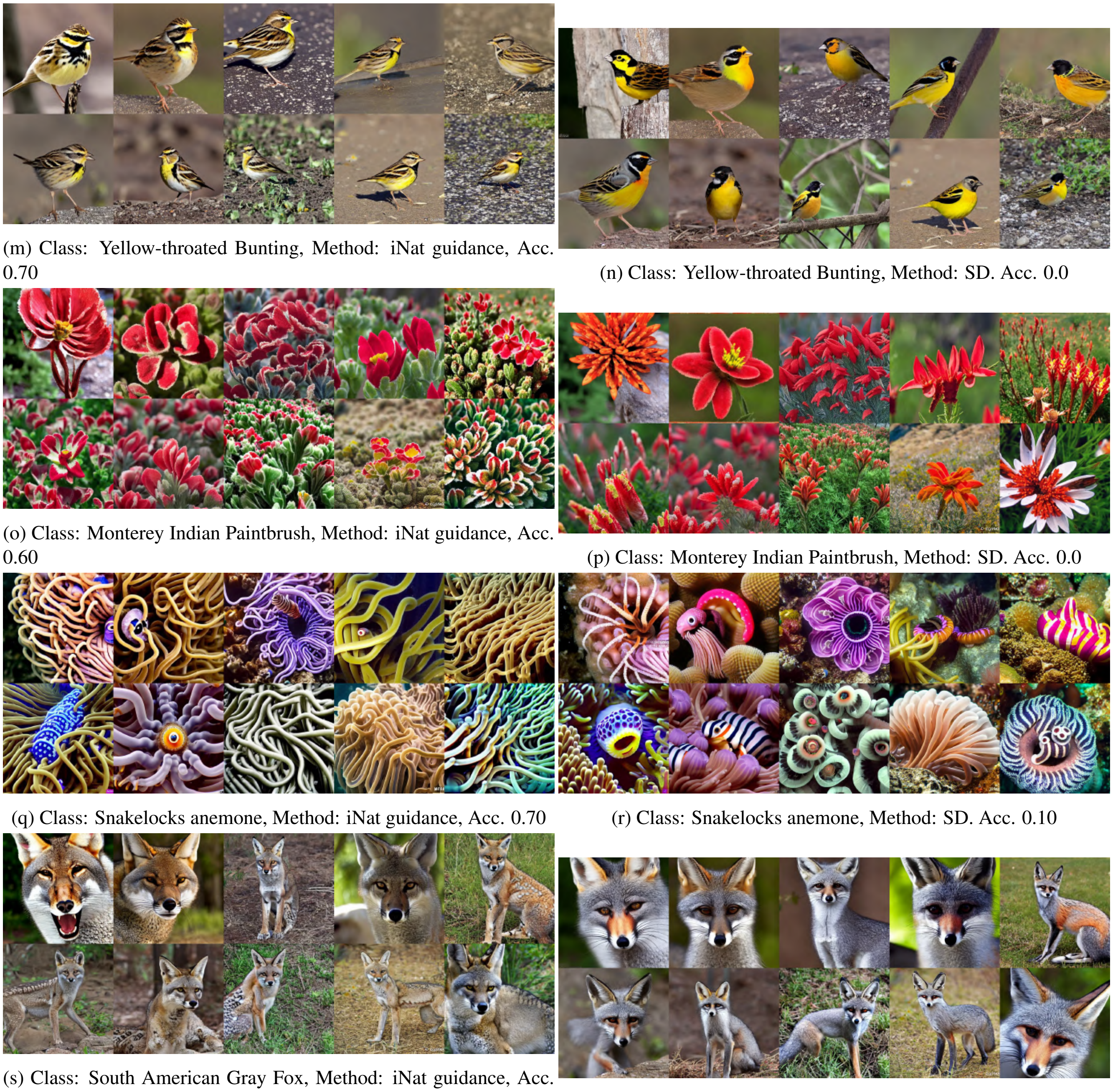
(h) Class: Red-bellied Squirrel, Method: SD. Acc. 0.0

(i) Class: Tricolored Bat, Method: iNat guidance, Acc. 0.70

(j) Class: Tricolored Bat, Method: SD. Acc. 0.0

(k) Class: Yellow-pine Chipmunk, Method: iNat guidance, Acc. 0.70

(l) Class: Yellow-pine Chipmunk, Method: SD. Acc. 0.0



(m) Class: Yellow-throated Bunting, Method: iNat guidance, Acc. 0.70

(n) Class: Yellow-throated Bunting, Method: SD. Acc. 0.0

(o) Class: Monterey Indian Paintbrush, Method: iNat guidance, Acc. 0.60

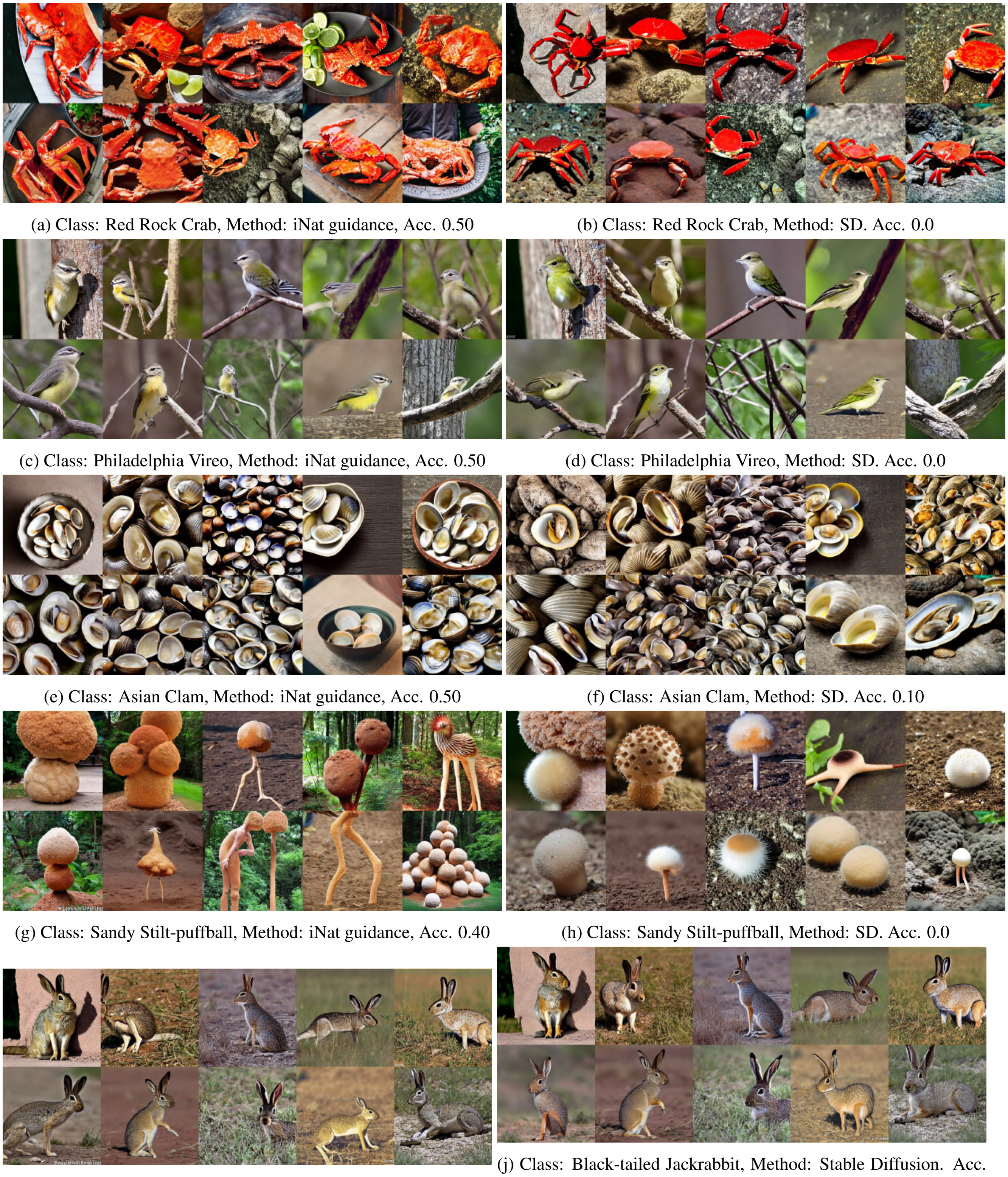
(p) Class: Monterey Indian Paintbrush, Method: SD. Acc. 0.0

(q) Class: Snakelocks anemone, Method: iNat guidance, Acc. 0.70

(r) Class: Snakelocks anemone, Method: SD. Acc. 0.10

(s) Class: South American Gray Fox, Method: iNat guidance, Acc. 1.00

(t) Class: South American Gray Fox, Method: SD. Acc. 0.40



(a) Class: Red Rock Crab, Method: iNat guidance, Acc. 0.50

(b) Class: Red Rock Crab, Method: SD. Acc. 0.0

(c) Class: Philadelphia Vireo, Method: iNat guidance, Acc. 0.50

(d) Class: Philadelphia Vireo, Method: SD. Acc. 0.0

(e) Class: Asian Clam, Method: iNat guidance, Acc. 0.50

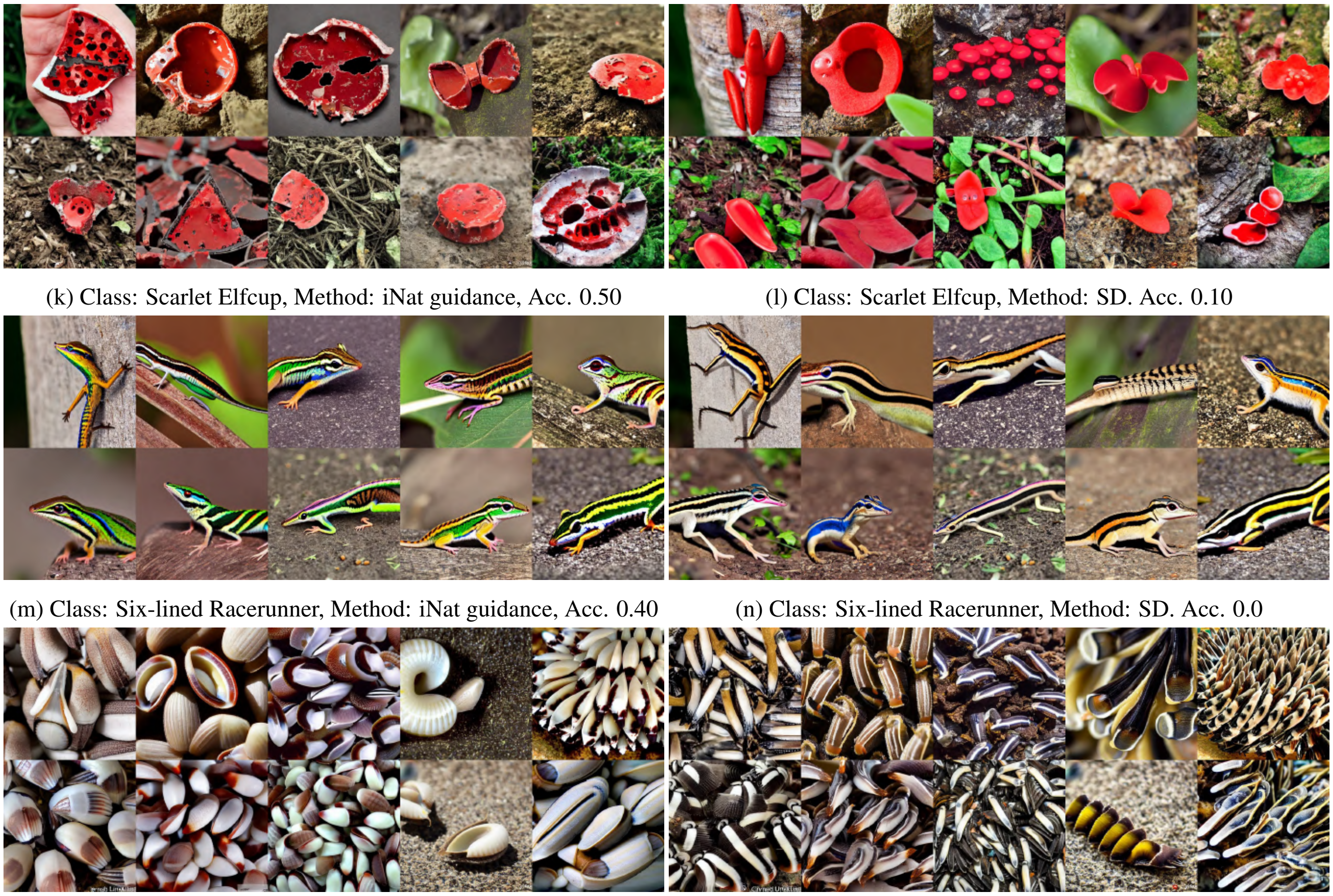
(f) Class: Asian Clam, Method: SD. Acc. 0.10

(g) Class: Sandy Stilt-puffball, Method: iNat guidance, Acc. 0.40

(h) Class: Sandy Stilt-puffball, Method: SD. Acc. 0.0

(i) Class: Black-tailed Jackrabbit, Method: iNat guidance, Acc. 0.7

(j) Class: Black-tailed Jackrabbit, Method: Stable Diffusion. Acc. 0.0



(k) Class: Scarlet Elfcup, Method: iNat guidance, Acc. 0.50

(l) Class: Scarlet Elfcup, Method: SD. Acc. 0.10

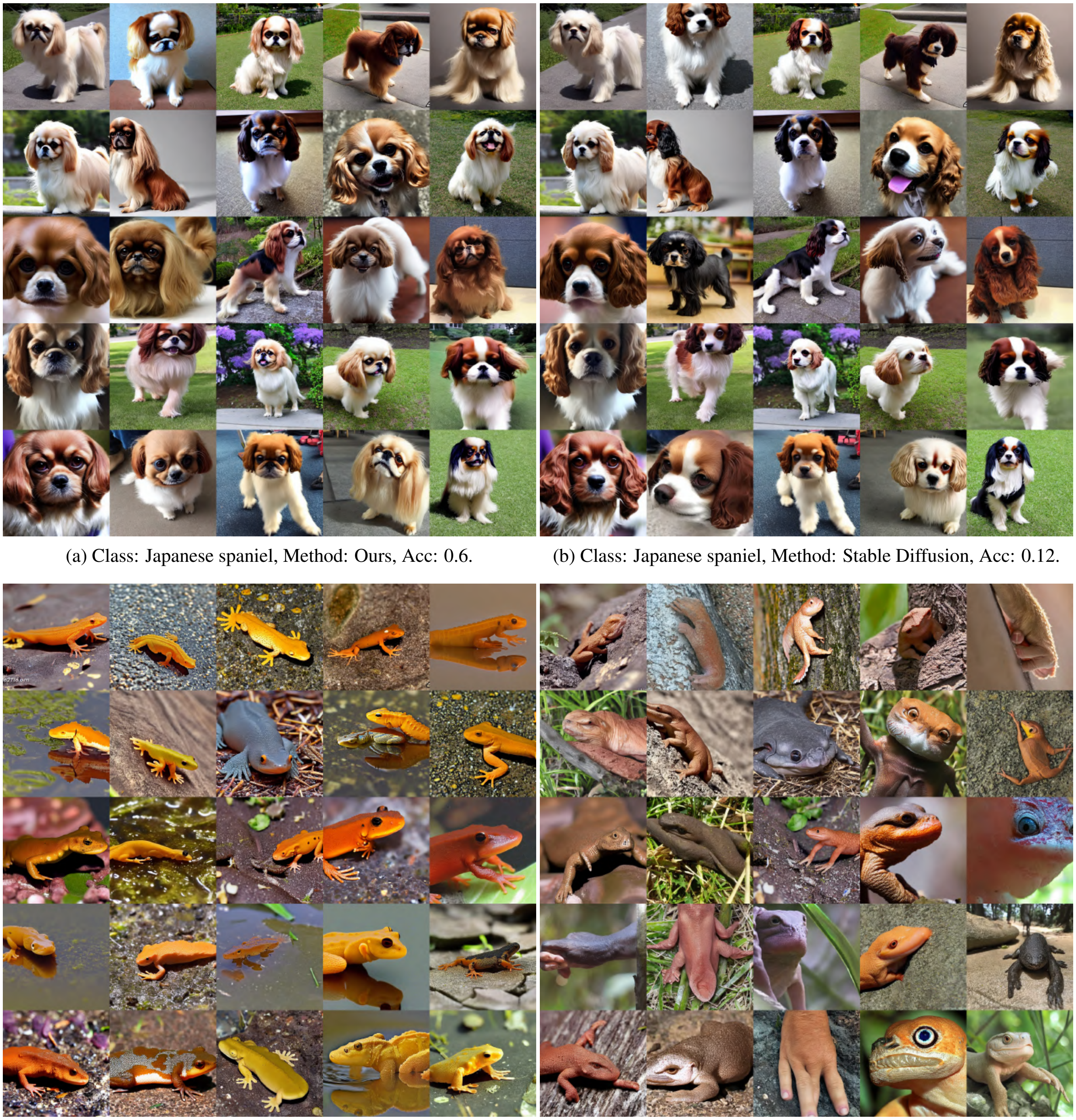
(m) Class: Six-lined Racerunner, Method: iNat guidance, Acc. 0.40

(n) Class: Six-lined Racerunner, Method: SD. Acc. 0.0

(o) Class: Goose Barnacle, Method: iNat guidance, Acc. 0.55

(p) Class: Goose Barnacle, Method: SD. Acc. 0.0

Figure 9: Images generated based on ImageNet classes. We show results with our discriminative token and vanilla stable diffusion. We provide the accuracy of the classification for each batch.

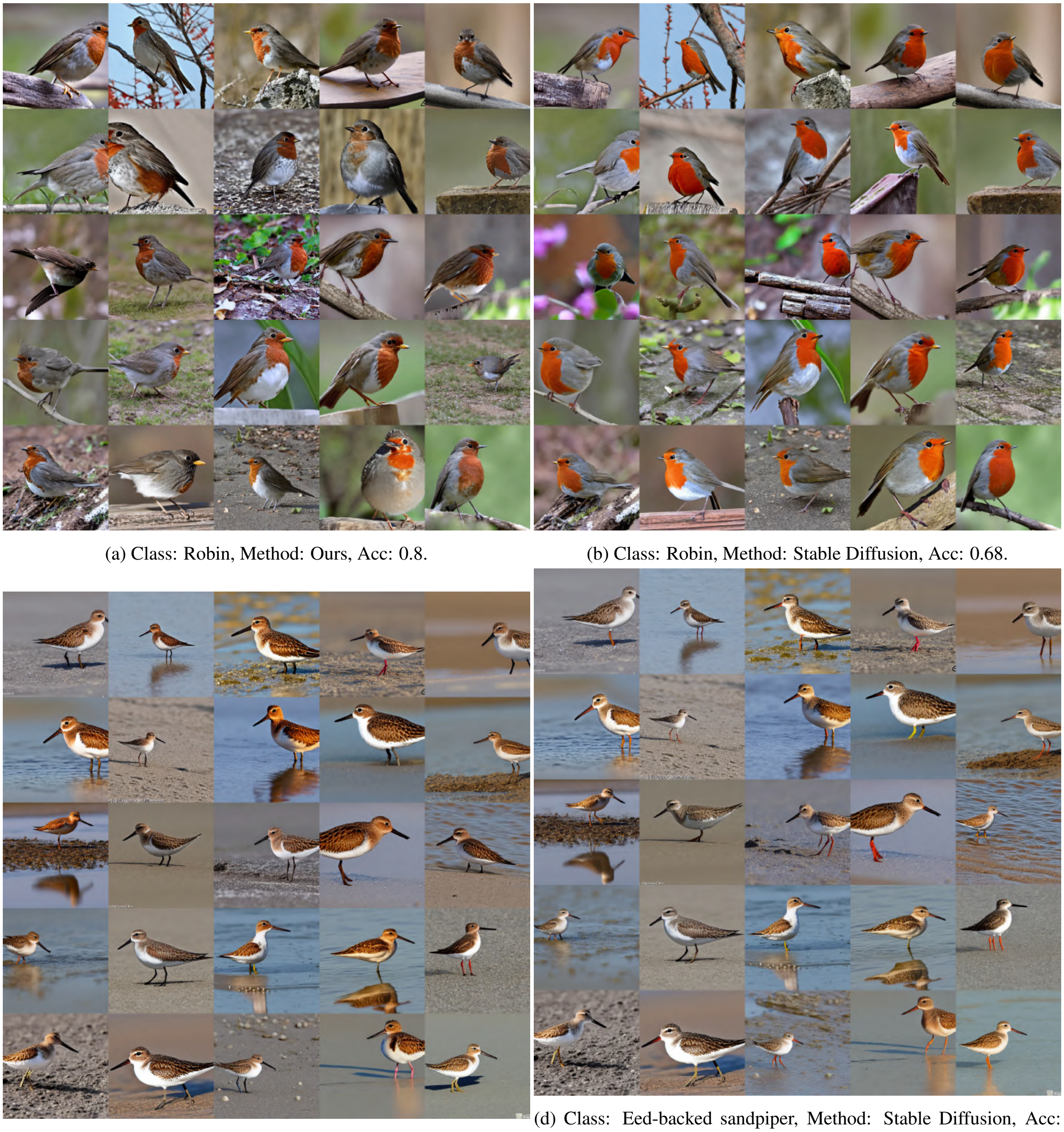


(a) Class: Japanese spaniel, Method: Ours, Acc: 0.6.

(b) Class: Japanese spaniel, Method: Stable Diffusion, Acc: 0.12.

(c) Class: eft, Method: Ours, Acc: 0.56.

(d) Class: eft, Method: Stable Diffusion, Acc: 0.06.

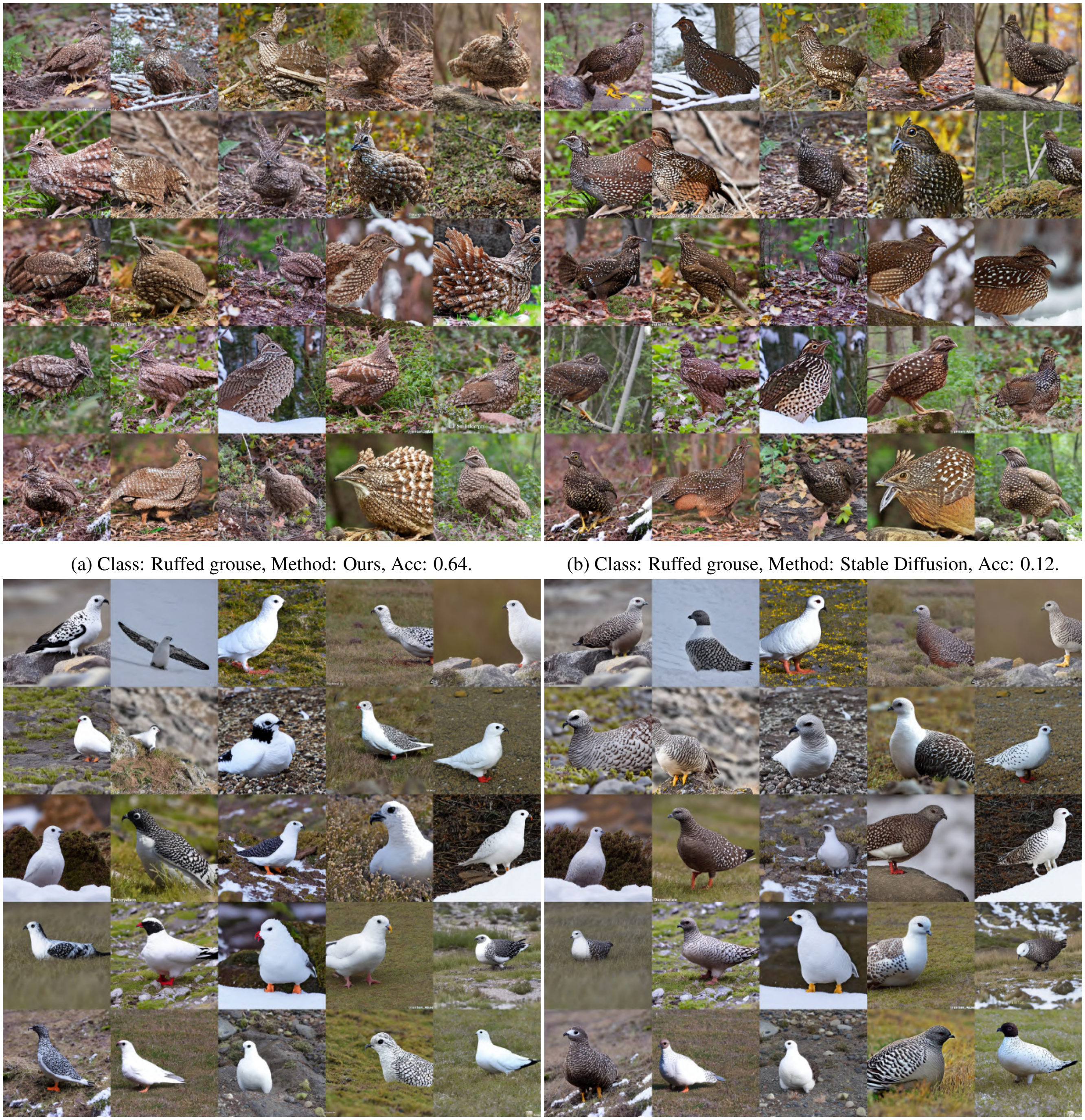


(a) Class: Robin, Method: Ours, Acc: 0.8.

(b) Class: Robin, Method: Stable Diffusion, Acc: 0.68.

(c) Class: Red-backed sandpiper, Method: Ours, Acc: 0.92.

(d) Class: Eed-backed sandpiper, Method: Stable Diffusion, Acc: 0.56.

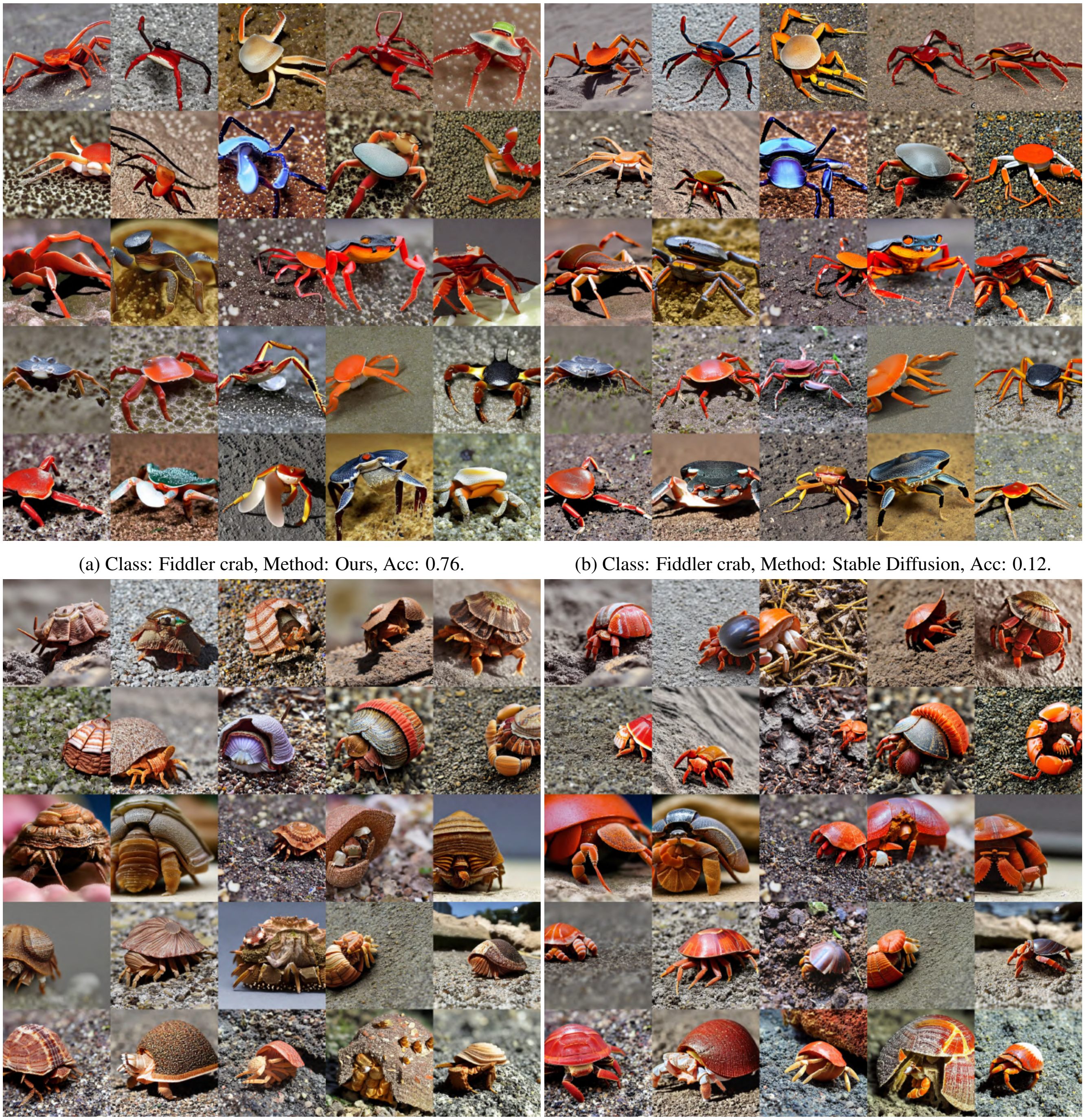


(a) Class: Ruffed grouse, Method: Ours, Acc: 0.64.

(b) Class: Ruffed grouse, Method: Stable Diffusion, Acc: 0.12.

(c) Class: Ptarmigan, Method: Ours, Acc: 0.64.

(d) Class: Ptarmigan, Method: Stable Diffusion, Acc: 0.08.

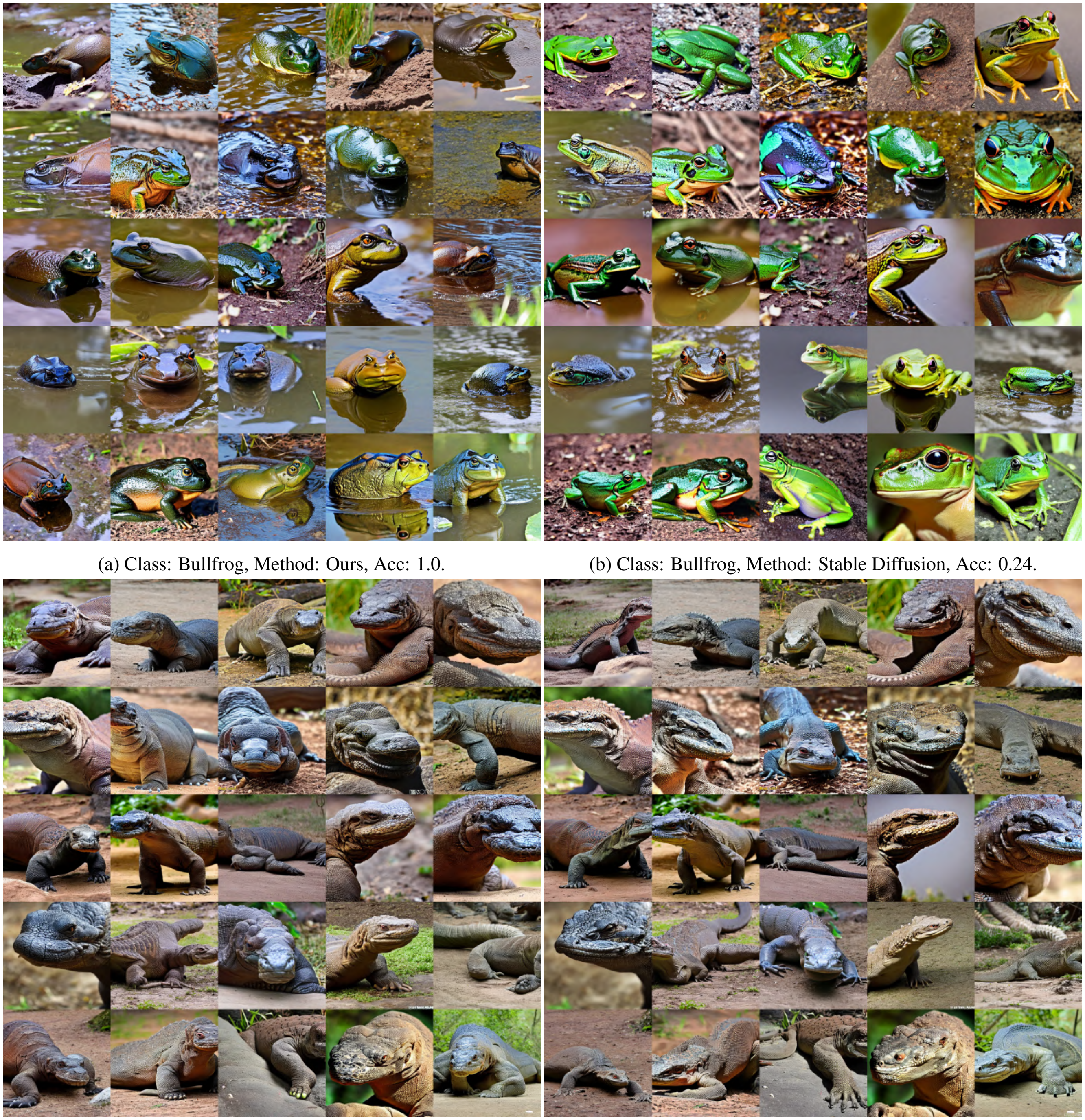


(a) Class: Fiddler crab, Method: Ours, Acc: 0.76.

(b) Class: Fiddler crab, Method: Stable Diffusion, Acc: 0.12.

(c) Class: Hermit crab, Method: Ours, Acc: 0.4.

(d) Class: Hermit crab, Method: Stable Diffusion, Acc: 0.12.

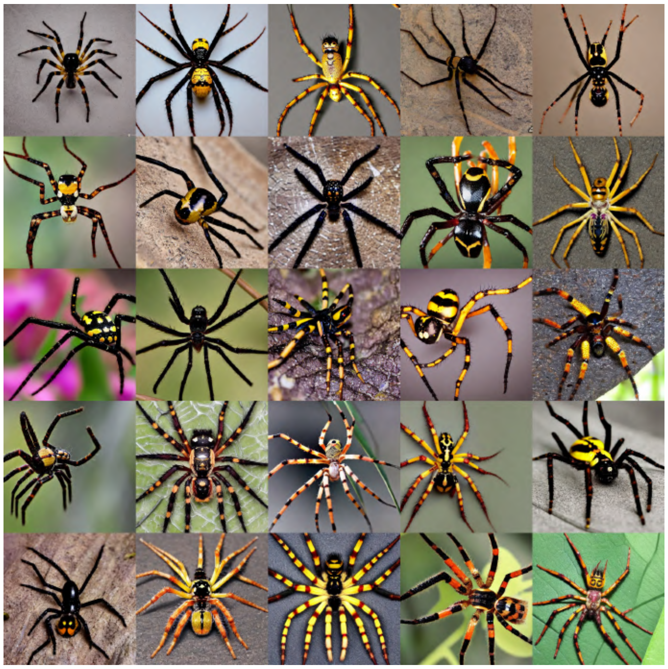


(a) Class: Bullfrog, Method: Ours, Acc: 1.0.

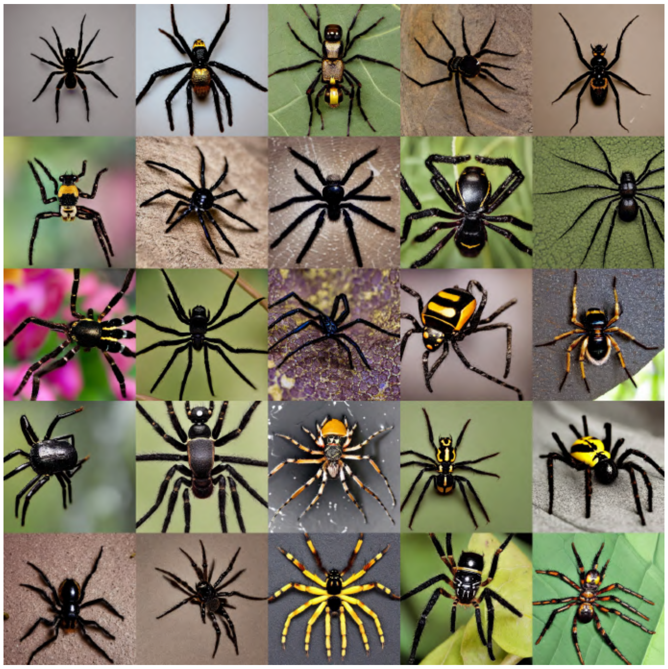
(b) Class: Bullfrog, Method: Stable Diffusion, Acc: 0.24.

(c) Class: Komodo dragon, Method: Ours, Acc: 0.92.

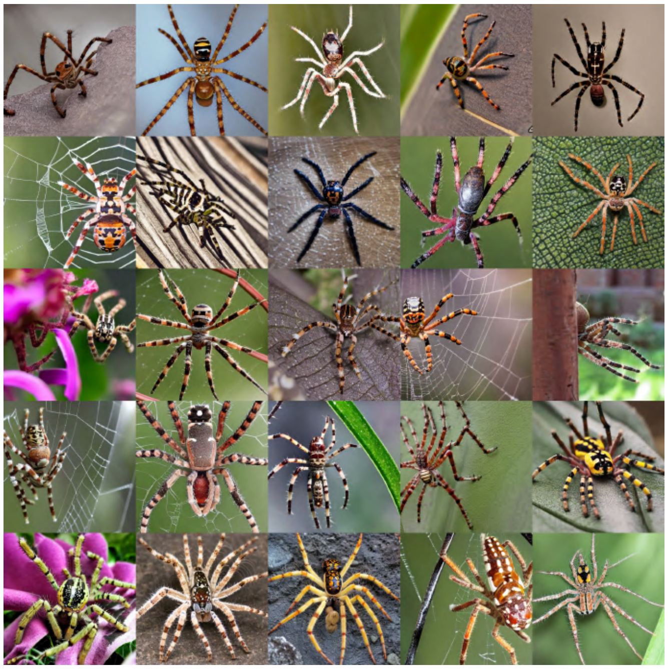
(d) Class: Komodo dragon, Method: Stable Diffusion, Acc: 0.48.



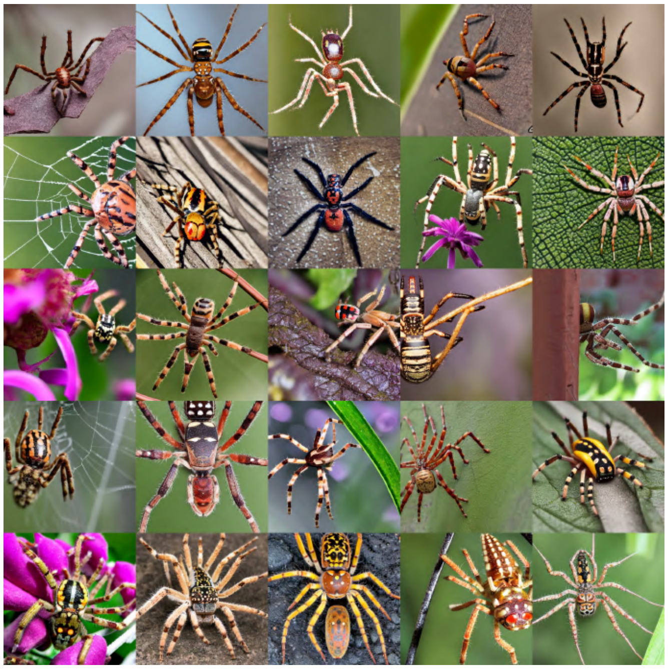
(a) Class: black and gold garden spider, Method: Ours, Acc: 0.16.



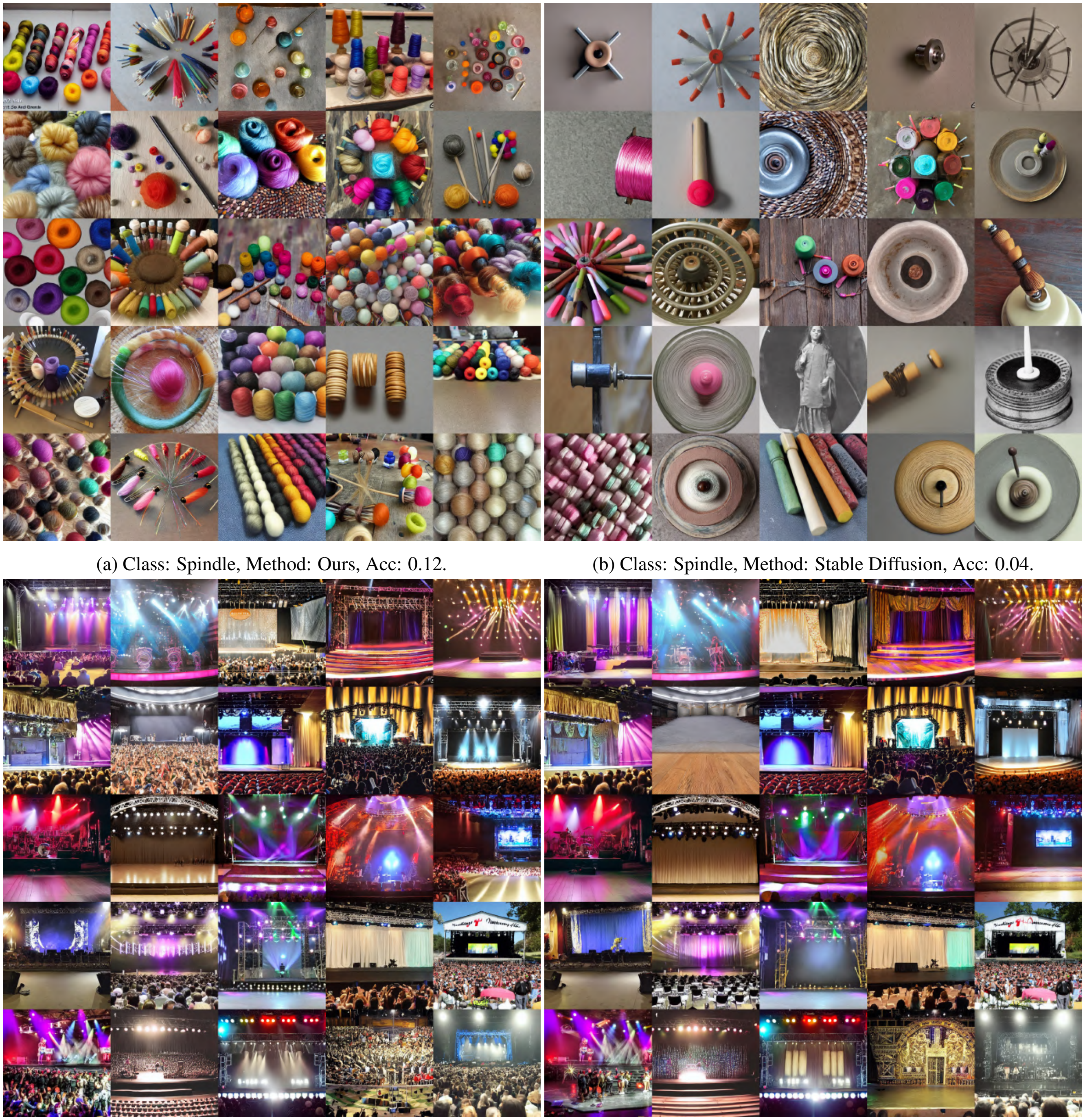
(b) Class: Black and gold garden spider, Method: Stable Diffusion, Acc: 0.0.



(c) Class: Garden spider, Method: Ours, Acc: 0.4.



(d) Class: Garden spider, Method: Stable Diffusion, Acc: 0.28.

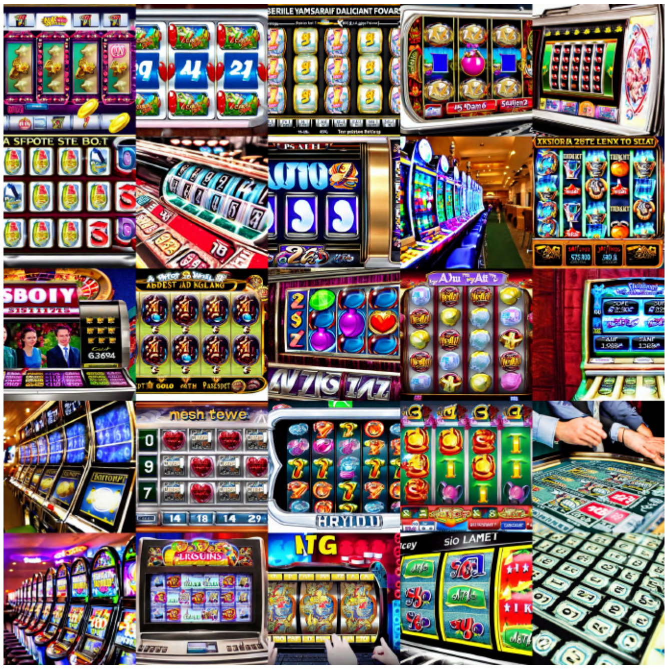


(a) Class: Spindle, Method: Ours, Acc: 0.12.

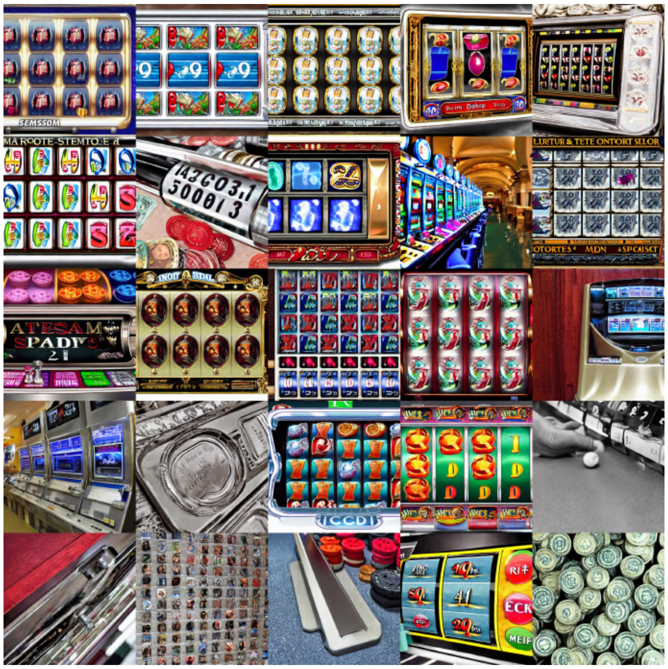
(b) Class: Spindle, Method: Stable Diffusion, Acc: 0.04.

(c) Class: Stage, Method: Ours, Acc: 0.88.

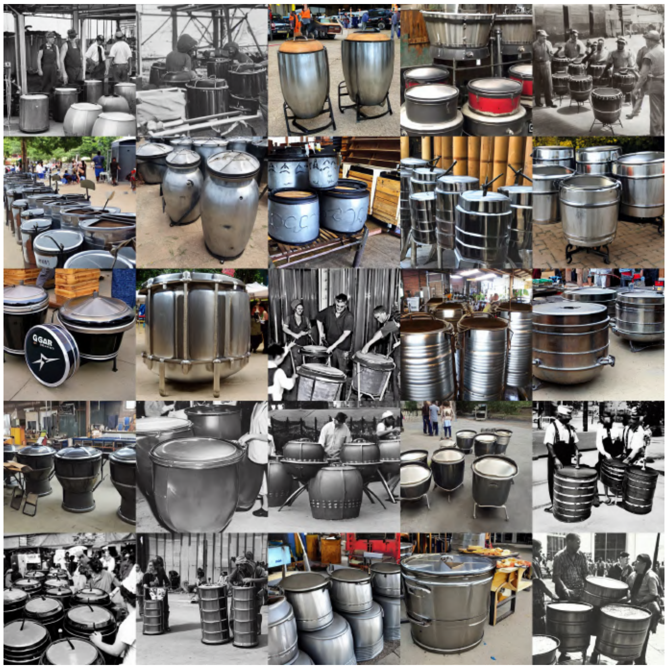
(d) Class: Stage, Method: Stable Diffusion, Acc: 0.48.



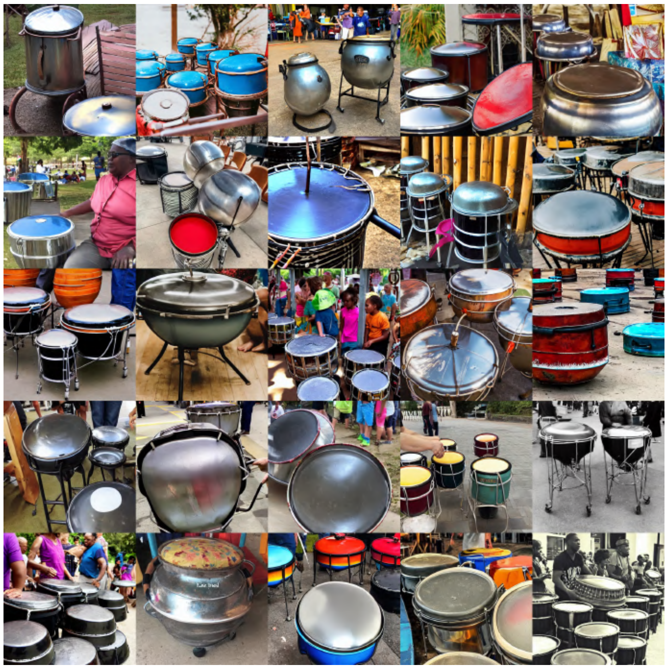
(a) Class: Slot, Method: Ours, Acc: 0.96.



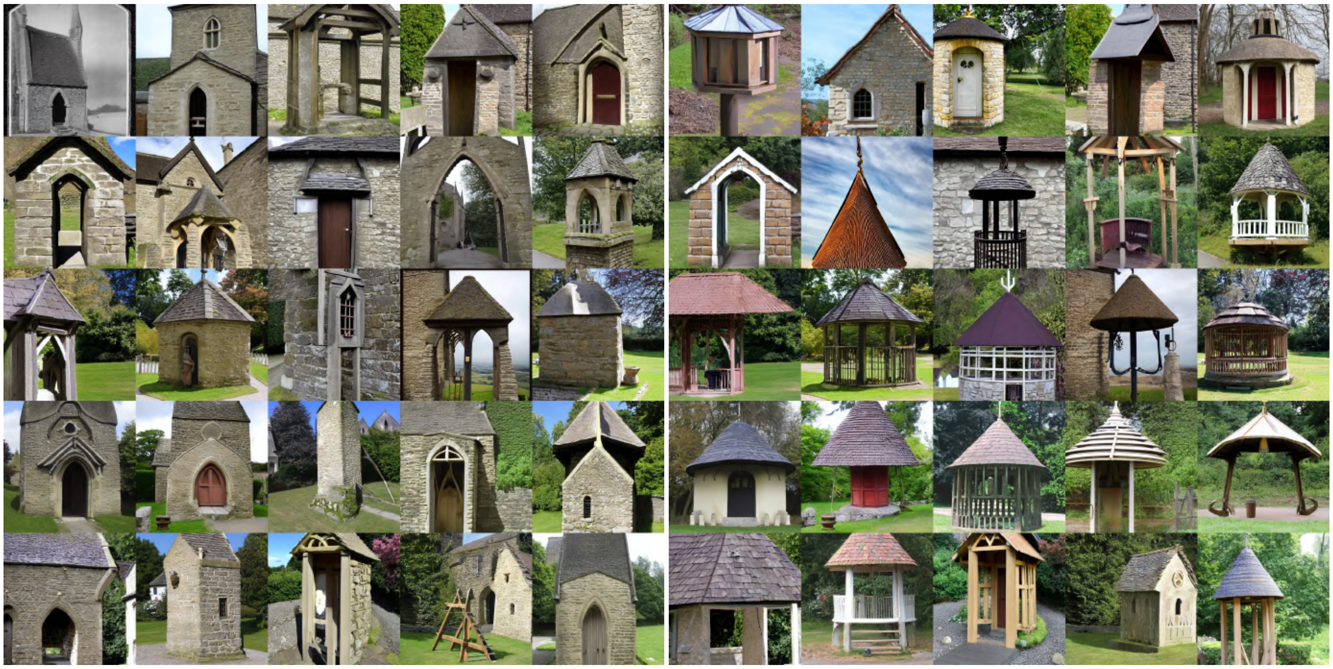
(b) Class: Slot, Method: Stable Diffusion, Acc: 0.64.



(c) Class: Steel drum, Method: Ours, Acc: 0.28.

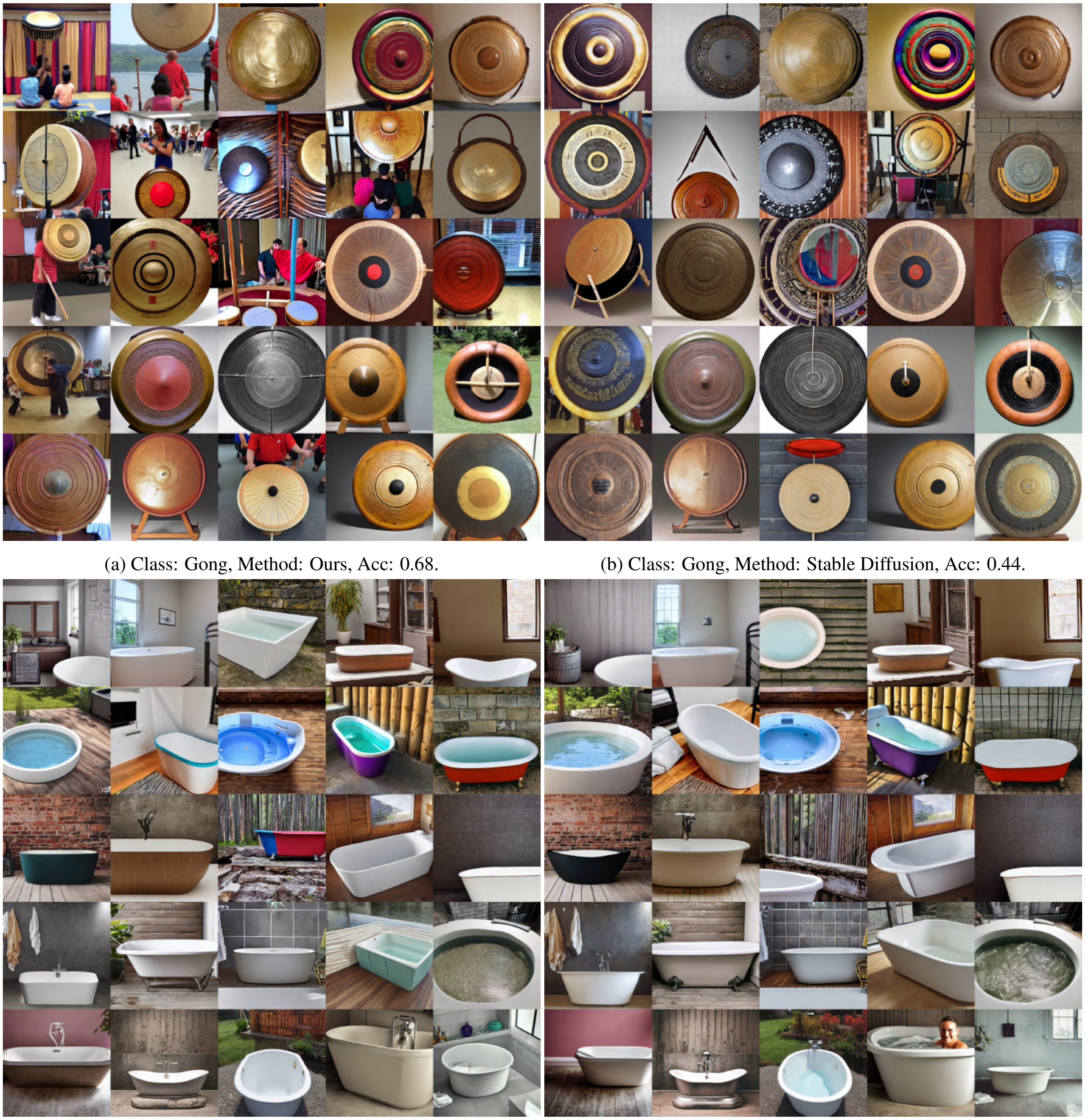


(d) Class: Steel drum, Method: Stable Diffusion, Acc: 0.16.



(a) Class: Bell cote, Method: Ours, Acc: 0.88.

(b) Class: Bell cote, Method: Stable Diffusion, Acc: 0.08.

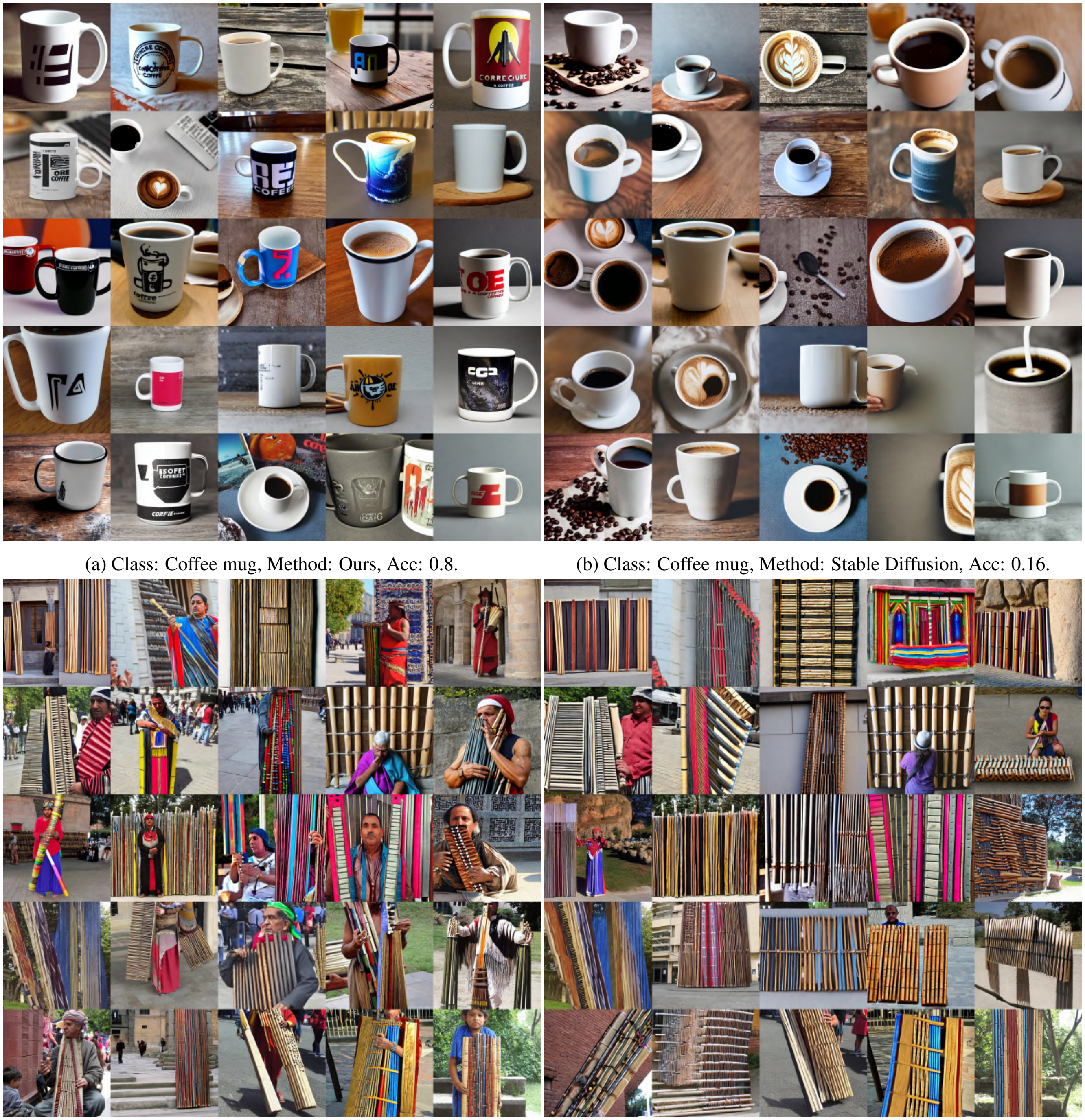


(a) Class: Gong, Method: Ours, Acc: 0.68.

(b) Class: Gong, Method: Stable Diffusion, Acc: 0.44.

(c) Class: Tub, Method: Ours, Acc: 0.72.

(d) Class: Tub, Method: Stable Diffusion, Acc: 0.48.



(a) Class: Coffee mug, Method: Ours, Acc: 0.8.

(b) Class: Coffee mug, Method: Stable Diffusion, Acc: 0.16.

(c) Class: Panpipe, Method: Ours, Acc: 0.36.

(d) Class: Panpipe, Method: Stable Diffusion, Acc: 0.08.

References

[1] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 10