IMS3: Breaking Distributional Aggregation in Diffusion-Based Dataset Distillation

Supplementary Material

Contents

7. Instability in Inversion

Diffusion inversion theoretically provides a deterministic mapping between real data and their latent noises by integrating the probability flow ODE (PF-ODE) backward in time. However, recent work reveals that this inversion process is inherently unstable in high-dimensional spaces [46]. Specifically, even infinitesimal perturbations in the latent space can be amplified through the ODE dynamics, leading to substantial reconstruction errors during the forward regeneration process.

Formally, let $F : \mathbb{R}^n \to \mathbb{R}^n$ denote the PF-ODE mapping from noise to data. The instability of F can be characterized by the geometric mean of its singular values, known as the *geometric mean instability coefficient*:

$$\bar{\mathcal{E}}_F(\mathbf{z}) = \left(\prod_{i=1}^{n} \frac{\|J_F(\mathbf{z})\,\mathbf{u}_i\|_2}{\|\mathbf{u}_i\|_2} \right)^{1/n}, \tag{12}$$

where $J_F(\mathbf{z})$ denotes the Jacobian matrix of F at $\mathbf{z}$ and $\{\mathbf{u}_i\}$ are orthonormal basis vectors. When $\bar{\mathcal{E}}_F(\mathbf{z}) > 1$, the mapping locally expands the volume element around $\mathbf{z}$, implying that infinitesimal perturbations will be amplified after propagation through F.

To quantify the probability that such instability occurs, they establish the following lower bound for any threshold $M > 1$:

$$\mathcal{P}_M := \pi_{\text{real}}\left(\{\mathbf{z} : \bar{\mathcal{E}}_F(F^{-1}(\mathbf{z})) > M\}\right) \geq 1 - \epsilon - \delta, \tag{13}$$

where

$$\begin{aligned} \epsilon &:= \pi_{\text{real}}\left(\left\{\mathbf{z} : p_{\text{gen}}(\mathbf{z}) \geq \frac{1}{(2\pi M^2)^{n/2}} e^{-\frac{2n+3\sqrt{2n}}{2}}\right\}\right), \\ \delta &:= \pi_{\text{real}}\left(\left\{\mathbf{z} : \|F^{-1}(\mathbf{z})\|^2 > 2n + 3\sqrt{2n}\right\}\right). \end{aligned} \tag{14}$$

Here, π_{real} denotes the real data distribution, and p_{gen} represents the generation distribution of the diffusion model. As dimensionality n increases, both ϵ and δ vanish, implying that the instability probability $\mathcal{P}_M$ approaches one—thus instability becomes almost inevitable in high-dimensional diffusion mappings.

Intuitively, this phenomenon arises from the sparsity of the generation distribution. Since the generative probability mass is concentrated in a few narrow regions while most of the space exhibits near-zero density, the PF-ODE must stretch the mapping to preserve probability. Consequently, regions of low density in the generative space correspond to large local Jacobian norms, leading to the amplification of any perturbation during inversion or regeneration. This explains why inversion trajectories often drift toward low-density, high-sensitivity regions of the data manifold, as also observed in Sec. 4.1. A more detailed mathematical derivation and discussion of the instability mechanism can be found in [46].

This inherent instability directly motivates the design of our Inversion-Matching fine-tuning (`IM`) in ImS^3. Rather than viewing instability as an obstacle, we exploit it as a guiding property to drive the diffusion model toward low-density regions.

By aligning the denoising trajectory with the inversion trajectory, our method explicitly leverages the instability-induced drift to expand the distributional coverage and enhance representation diversity in the distilled dataset.

8. More Experiments

8.1. Comparison on ImageNet-100

Tab. 5 reports the comparison across different IPC and backbone on ImageNet-100. Across all settings, our method consistently delivers the highest accuracy, outperforming existing approaches such as Herding [41], IDC-1 [14], Minimax [8], and MGD^3 [3].

Table 5. Performance comparison on ImageNet-100 under different IPC and backbone settings. Results are Top-1 accuracy.

IPC	Model	Random	Herding [41]	IDC-1 [14]	Minimax [14]	MGD^3 [3]	ImS^3 (Ours)
10	ConvNet-6	17.0$_{\pm 0.3}$	17.2$_{\pm 0.3}$	24.3$_{\pm 0.5}$	22.3$_{\pm 0.5}$	23.4$_{\pm 0.9}$	**24.3**$_{\pm 1.1}$
	ResNetAP-10	19.1$_{\pm 0.4}$	19.8$_{\pm 0.3}$	25.7$_{\pm 0.1}$	24.8$_{\pm 0.2}$	25.8$_{\pm 0.5}$	**28.4**$_{\pm 0.2}$
	ResNet-18	17.5$_{\pm 0.5}$	16.1$_{\pm 0.2}$	25.1$_{\pm 0.2}$	22.5$_{\pm 0.3}$	23.6$_{\pm 0.4}$	**28.3**$_{\pm 0.3}$
20	ConvNet-6	24.8$_{\pm 0.2}$	24.3$_{\pm 0.4}$	28.8$_{\pm 0.3}$	29.3$_{\pm 0.4}$	30.6$_{\pm 0.4}$	**30.7**$_{\pm 0.4}$
	ResNetAP-10	26.7$_{\pm 0.5}$	27.6$_{\pm 0.1}$	29.9$_{\pm 0.2}$	32.3$_{\pm 0.1}$	33.9$_{\pm 1.1}$	**34.8**$_{\pm 0.2}$
	ResNet-18	25.5$_{\pm 0.3}$	24.7$_{\pm 0.1}$	30.2$_{\pm 0.2}$	31.2$_{\pm 0.1}$	32.6$_{\pm 0.4}$	**36.3**$_{\pm 0.2}$

8.2. Experiment Results on Different Student Model

Tab. 6 reports the performance of different dataset distillation methods on ImageWoof across multiple backbone architectures,including ResNet-18, ResNet-50, ResNet-101, and multiple IPC settings. Across all configurations, our method achieves the highest accuracy, demonstrating clear and consistent advantages in both low- and high-budget regimes. The improvements are particularly notable under low IPC (e.g., IPC = 10), where the limited data makes representation learning more challenging: ImS^3 boosts ResNet-18 performance from 45.6% (CaO^2) to 45.9%, and ResNet-50 accuracy from 40.1% to 42.7%. Under higher IPC (IPC = 50), ImS^3 continues to outperform strong baselines such as Minimax [8], RDED [33], and CaO^2 [37], confirming that our approach generalizes well across model capacities and data regimes. These results highlight the robustness and scalability of ImS^3 for improving distilled dataset quality on large-scale visual recognition tasks.

8.3. Analysis on Loss Functions

Table 7 presents an ablation study on different similarity losses used in IM. We evaluate three widely used losses L_1, L_2, and $1-\sigma$ in Sec. 4.1 on ImageWoof and ImageNette under IPC = 10 and 50. Both L_1 and L_2 exhibit comparable performance, providing moderate improvements across IPC settings. In contrast, the proposed $1-\sigma$ formulation consistently achieves the highest accuracy, surpassing the other two losses by a clear margin at both data budgets. These results highlight the importance of treating feature deviations in a distribution-aware manner, suggesting that the $1-\sigma$-based loss stabilizes inversion alignment and leads to more reliable distilled samples.

8.4. Analysis on λ_{IM}

To examine the sensitivity of our method to the matching strength, we perform a sweep over λ_{IM} and report the validation accuracy for IPC = 10, 20, and 50 in Fig. 5). Across all settings, the performance remains stable within a broad interval of λ_{IM}, indicating that the method is not overly sensitive to this hyperparameter. For smaller coefficients (*e.g.,* 2×10^{-3} to 8×10^{-3}), the accuracy stays consistently high, especially at IPC = 50. When λ_{IM} becomes excessively large (*e.g.,* 0.4–0.8), the accuracy gradually drops, likely because overly strong alignment introduces optimization bias and reduces the diversity of the distilled samples.

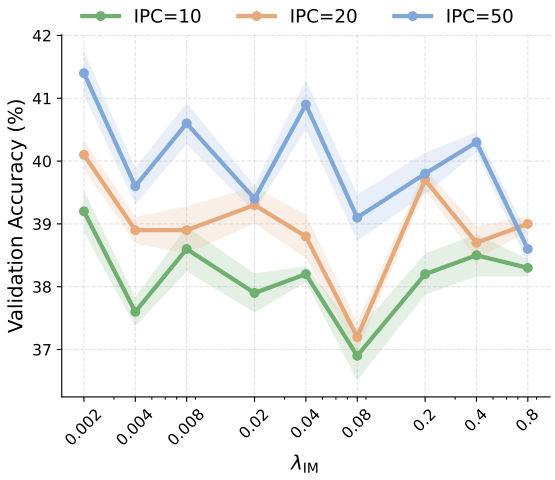


Figure 5. Validation accuracy under different matching strengths. Performance of our method across a range of matching coefficients λ_{IM} for IPC = 10, 20, and 50. Each curve corresponds to a fixed IPC, and shaded regions denote performance variability across runs.

8.5. Analysis on G

We analyze the effect of the subgroup pool size G, which determines how many candidate subgroups are generated for each class before selection. As shown in Fig. 6, increasing G helps expose more inter-class variability and improves distilled accuracy when the IPC is small, since low-data regimes benefit from a richer candidate pool. However, the trend does not extend indefinitely: overly large pools introduce redundant or low-quality samples, weakening the discriminative structure and causing performance to decline. This highlights the need to choose G according to the IPC setting, where lower IPC typically requires a larger

Table 6. Performance comparison under different IPC and backbone settings on ImageWoof. Best results in each setting are marked as bold.

Backbone	IPC	SRe^{2}L [45]	Minimax [8]	RDED [33]	CaO^2 [37]	ImS^3 (Ours)
ResNet-18	10	20.2$_{\pm 0.2}$	40.1$_{\pm 1.0}$	38.5$_{\pm 2.1}$	45.6$_{\pm 1.4}$	**45.9**$_{\pm 1.3}$
	50	23.3$_{\pm 0.3}$	67.0$_{\pm 1.8}$	68.5$_{\pm 0.7}$	68.9$_{\pm 1.1}$	**71.2**$_{\pm 1.3}$
ResNet-50	10	17.3$_{\pm 1.7}$	37.3$_{\pm 1.1}$	29.9$_{\pm 2.2}$	40.1$_{\pm 0.1}$	**42.7**$_{\pm 1.8}$
	50	24.8$_{\pm 0.7}$	64.3$_{\pm 0.9}$	67.8$_{\pm 0.3}$	68.2$_{\pm 1.1}$	**68.3**$_{\pm 0.2}$
ResNet-101	10	17.7$_{\pm 0.9}$	34.2$_{\pm 1.7}$	31.3$_{\pm 1.3}$	36.5$_{\pm 1.4}$	**38.2**$_{\pm 1.6}$
	50	21.2$_{\pm 0.2}$	62.7$_{\pm 1.6}$	59.1$_{\pm 0.7}$	63.1$_{\pm 1.3}$	**66.1**$_{\pm 1.8}$

Table 7. Ablation study of different similarity losses used in the IM on **ImageWoof** and **ImageNette** under IPC = 10 and 50. Results are reported as top-1 accuracy.

Loss Type	ImageWoof		ImageNette	
	IPC = 10	IPC = 50	IPC = 10	IPC = 50
L_1	38.7$_{\pm 0.2}$	58.1$_{\pm 0.2}$	62.6$_{\pm 0.9}$	81.7$_{\pm 0.7}$
L_2	39.0$_{\pm 0.2}$	58.1$_{\pm 0.1}$	62.7$_{\pm 1.0}$	82.8$_{\pm 0.7}$
$1-\sigma$	**41.8**$_{\pm 0.3}$	**60.1**$_{\pm 0.7}$	**62.9**$_{\pm 1.2}$	**84.2**$_{\pm 1.0}$

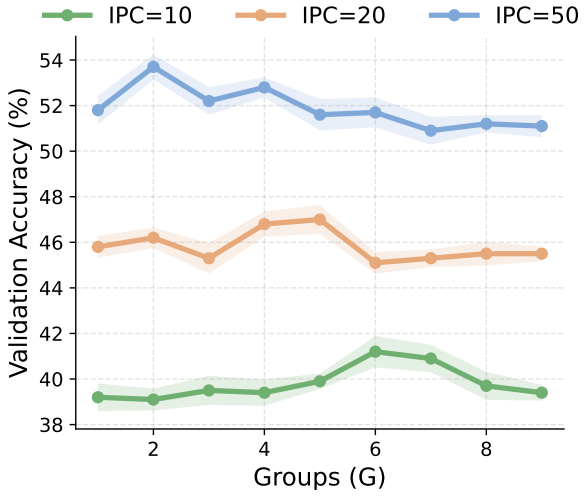


Figure 6. Effect of subgroup pool size G under different IPC settings on ImageWoof. A moderate increase in G provides richer intra-class variation and improves selection quality, especially in low-IPC regimes. However, excessively large pools introduce redundant or noisy candidates, which destabilizes selection and leads to degraded performance.

pool while higher IPC benefits less from further expansion.

8.6. Analysis on Feature Extractors for Centroid Selection

To further investigate the robustness of our centroid–based subgroup selection, we replace the default ResNet-18 feature extractor with different backbones, including ResNet-50, ResNet-101, EfficientNet, and CLIP. We report the distilled top-1 accuracy on ImageWoof under IPC = 10 and 20.

Table 8. Ablation study on feature extractors for centroid selection on ImageWoof. Top-1 accuracy (%).

IPC	Res18	Res50	Res101	Efficient	CLIP
10	**41.8**$_{\pm 0.3}$	39.6$_{\pm 0.3}$	38.8$_{\pm 0.2}$	39.5$_{\pm 0.2}$	38.1$_{\pm 0.5}$
20	**45.8**$_{\pm 1.2}$	45.7$_{\pm 0.3}$	45.0$_{\pm 0.2}$	45.8$_{\pm 0.3}$	44.3$_{\pm 0.2}$

Across both settings, we observe that the choice of feature extractor has noticable impact on the quality of sampling process. The results do not follow the conventional expectation that deeper backbones always yield better distilled results. Instead, the lightweight ResNet-18 achieves the highest top-1 accuracy among both IPC budgets. It implies that moderate capacity features provide a more stable and suitable embedding space for centroid sampling procedure.

8.7. t-SNE Visualization on ImageNet-100

To further examine how our method shapes the feature distribution of the distilled dataset, we visualize the embeddings of synthesized samples using t-SNE on ImageNet-100. We compare our approach (ImS^3) with MinimaxDiffusion, and the results are shown in Fig. 8. Each color corresponds to a different class.

As illustrated by the comparison, ImS^3 produces significantly more compact and well-structured intra-class clusters. Points belonging to the same class exhibit tighter grouping, indicating that our inversion-matching mechanism encourages stronger within-class consistency. More-

over, ImS^3 demonstrates clearer separation between different classes, suggesting that the synthesized samples form a more discriminative embedding structure. This indicates that our distribution-aware feature alignment promotes both intra-class cohesion and inter-class separability, which are crucial properties for effective dataset distillation.

In contrast, MinimaxDiffusion yields more fragmented and overlapping clusters. Classes become less compact and boundaries are looser, implying weaker semantic organization in the distilled data. The tighter clustering and improved class separation produced by ImS^3 provide intuitive evidence that our method captures the intrinsic class geometry more faithfully, leading to improved downstream recognition performance.

8.8. Runtime Analysis

We report runtime on a single RTX 4090 GPU using ImageWoof. As shown in Table 9, IM requires only 8 epochs of fine-tuning, while S^3 sampling is training-free, resulting in competitive total runtime.

Table 9. Runtime comparison on ImageWoof (IPC=10).

	DiT	Minimax	MGD3	**IMS^3**
Fine-tuning	-	32min	-	**19.5min**
Sampling	1.5min	**1.5min**	21min	7min
Total	1.5min	33.5min	**21min**	26.5min

8.9. Higher Resolution (512×512).

We evaluate IMS^3 at 512×512 using DiT-XL/2-512. Table 10 reports consistent advantages at IPC=10 and IPC=50.

Table 10. 512×512 results on ImageWoof.

	DiT	Minimax	CaO^2	**IMS^3**
IPC=10	36.6$_{\pm 0.2}$	33.3$_{\pm 0.2}$	37.5$_{\pm 0.3}$	**37.9**$_{\pm 0.2}$
IPC=50	52.5$_{\pm 0.2}$	51.4$_{\pm 0.5}$	53.8$_{\pm 0.6}$	**59.6**$_{\pm 0.3}$

8.10. ViT Architecture Evaluation.

We evaluate distilled datasets on ViT and EfficientNet. As shown in Table 11, IMS^3 improves results under both architectures, indicating good transfer across classifier families.

8.11. ImageNet-1K Experiments.

We evaluate IMS^3 on full ImageNet-1K. Table 12 reports results.

8.12. Decision Boundary Visualization.

We visualize classifier features using t-SNE in Figure 7. Samples selected by IMS^3 are closer to the decision boundaries.

Table 11. Cross-architecture evaluation on ImageWoof.

	Model	Minimax	MGD3	**IMS^3**
IPC=10	EfficientNet	31.3$_{\pm 0.7}$	32.3$_{\pm 0.3}$	**34.2**$_{\pm 0.2}$
	ViT	18.9$_{\pm 0.0}$	17.7$_{\pm 0.2}$	**20.0**$_{\pm 0.0}$
IPC=50	EfficientNet	46.3$_{\pm 1.0}$	49.3$_{\pm 0.5}$	**51.7**$_{\pm 0.4}$
	ViT	18.8$_{\pm 0.3}$	17.6$_{\pm 0.1}$	**19.9**$_{\pm 0.5}$

Table 12. ImageNet-1K results (IPC=10).

	SRe^{2}L	RDED	Minimax	MGD3	**IMS^3**
Top-1 Acc.	21.3$_{\pm 0.6}$	42.0$_{\pm 0.1}$	44.3$_{\pm 0.5}$	45.5$_{\pm 0.1}$	**45.6**$_{\pm 0.3}$

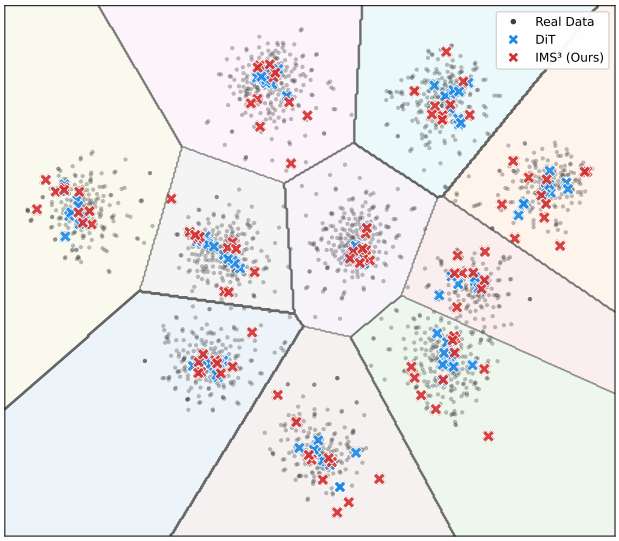


Figure 7. Feature-space visualization of distilled samples.

8.13. Real Data Accessibility

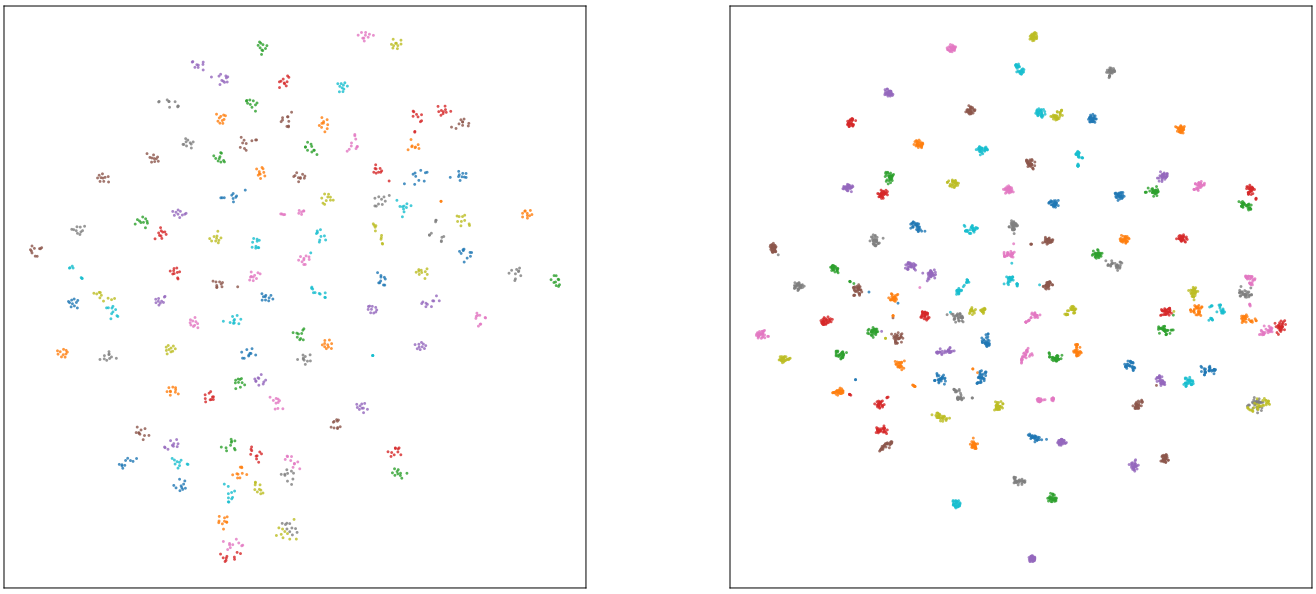
S³ can approximate class centroids using samples generated by the diffusion model itself. As shown in Table 13, using generated samples to estimate centroids still effective compared to the baselines.

Table 13. Gen-data centroid results on ImageWoof

	DiT	Minimax	**IMS$^3_{\text{Gen}}$**	**IMS$^3_{\text{Real}}$**
IPC=10	34.7$_{\pm 0.5}$	35.7$_{\pm 0.3}$	39.6$_{\pm 0.4}$	**41.8**$_{\pm 0.3}$
IPC=50	49.3$_{\pm 0.2}$	54.4$_{\pm 0.6}$	55.5$_{\pm 0.3}$	**57.3**$_{\pm 0.5}$

9. Visualization

We demonstrate the samples selected by Minimax [8] and ImS^3, Fig. 9 is the distilled samples from ImageWoof, Fig. 10 is the distilled samples from ImageNette, and



(a) ImageNet-100: ImS^3 (ours)

(b) ImageNet-100: MinimaxDiffusion

Figure 8. t-SNE visualization of distilled samples on ImageNet-100 and ImageWoof. ImS^3 produces tighter intra-class clusters and clearer inter-class separation, whereas MinimaxDiffusion exhibits more dispersed and overlapping structures.

Figs. 11 to 20 are the distilled samples from ImageNet-100.

Figure 9. Comparison between samples selected by Minimax [8] (left) and generated by the proposed ImS^3(right) for ImageWoof. The class names are marked at the left of each row.

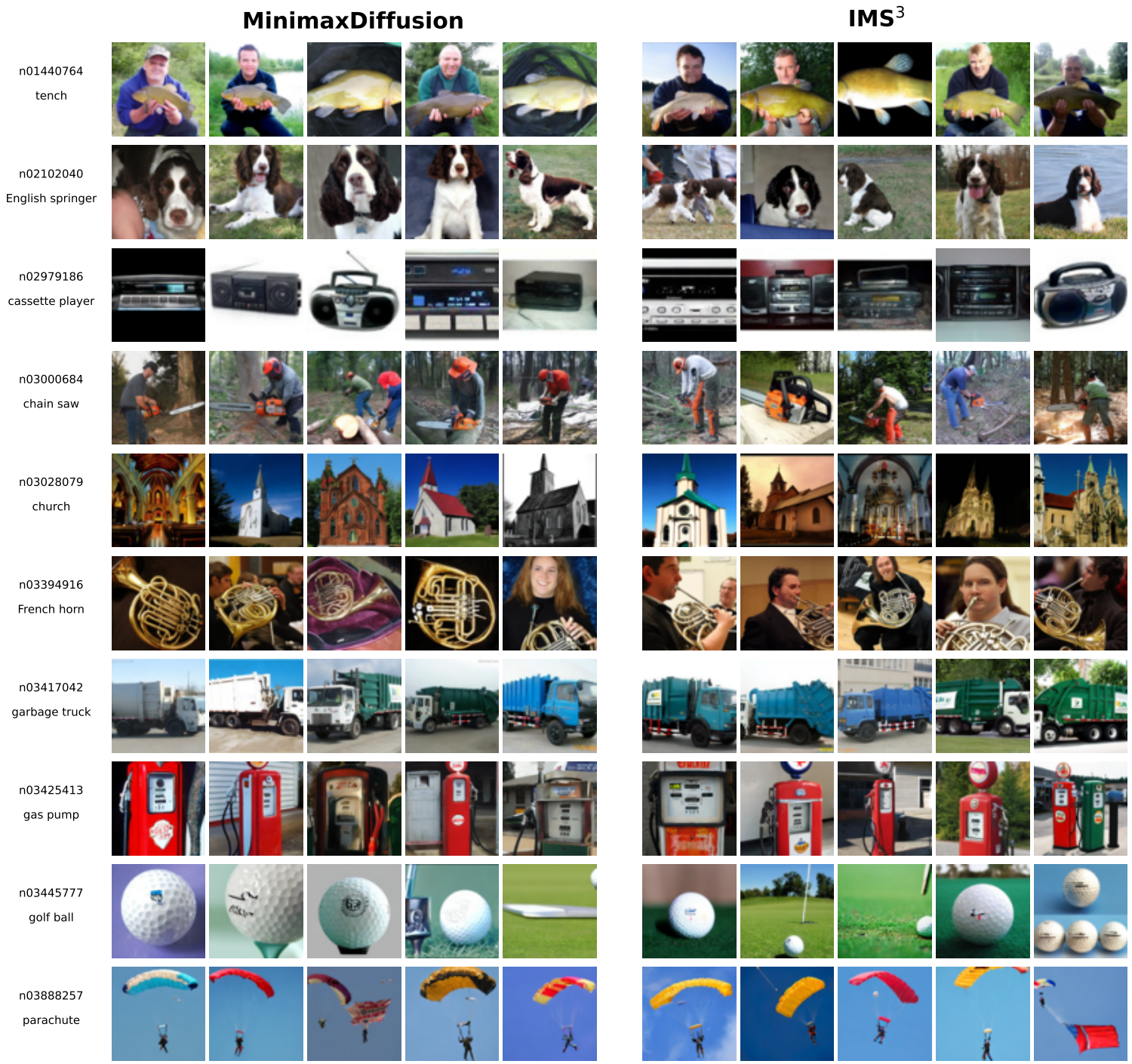


Figure 10. Comparison between samples selected by Minimax [8] (left) and generated by the proposed ImS^3(right) for ImageNette. The class names are marked at the left of each row.

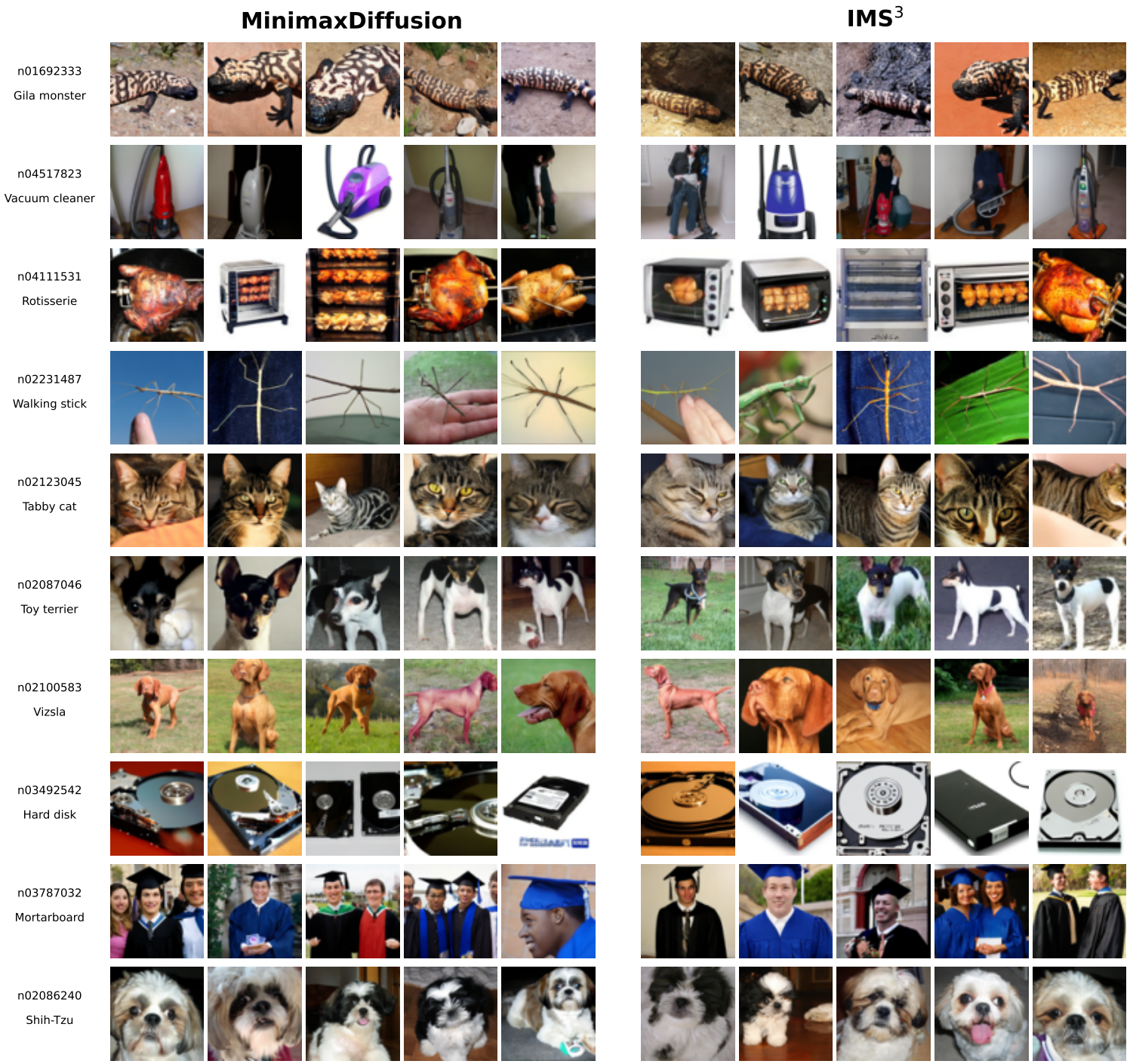


Figure 11. Comparison between samples selected by Minimax [8] (left) and generated by the proposed ImS^3(right) for ImageNet-100 classes 0-9. The class names are marked at the left of each row.

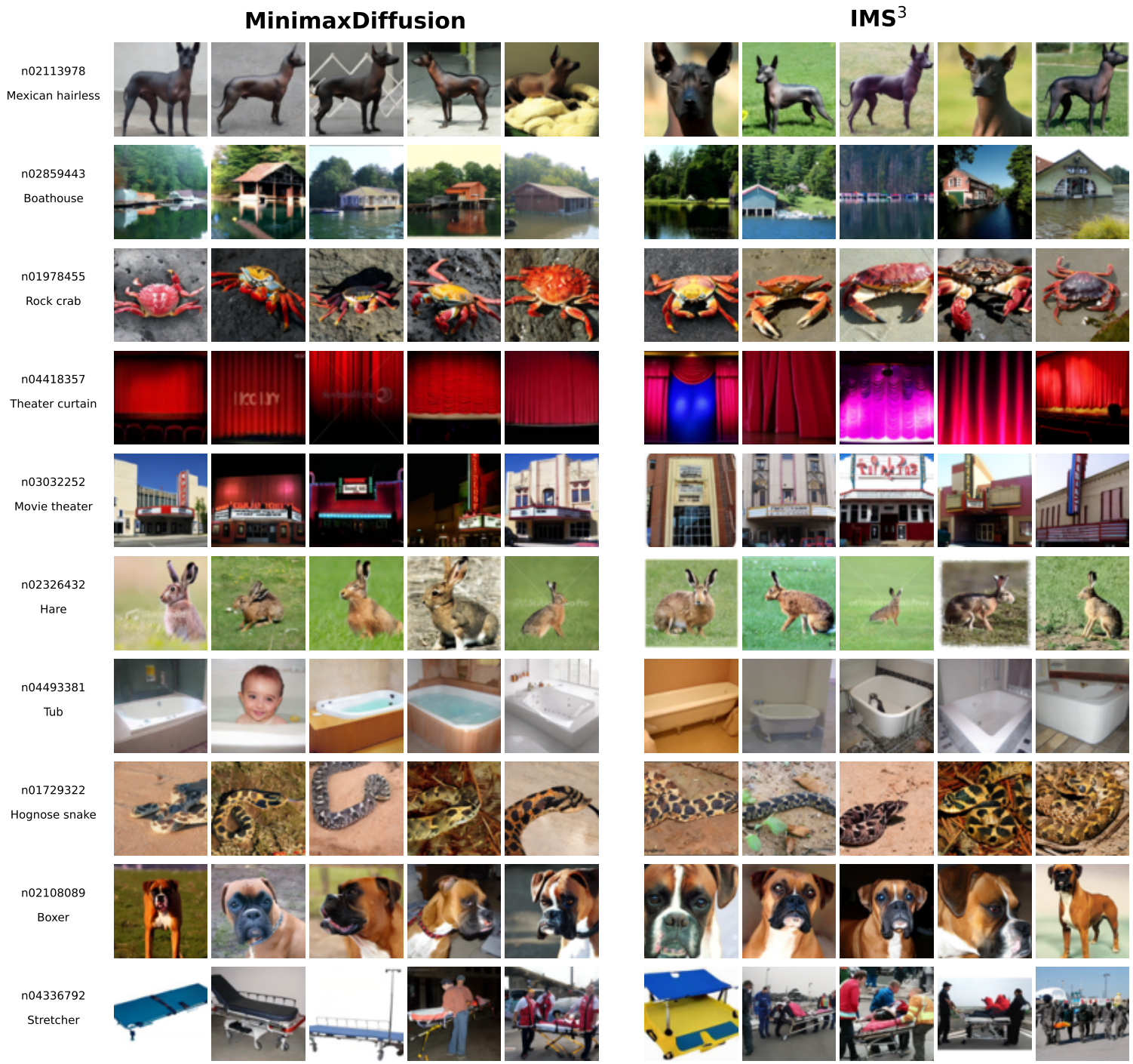


Figure 12. Comparison between samples selected by Minimax [8] (left) and generated by the proposed ImS^3(right) for ImageNet-100 classes 10-19. The class names are marked at the left of each row.

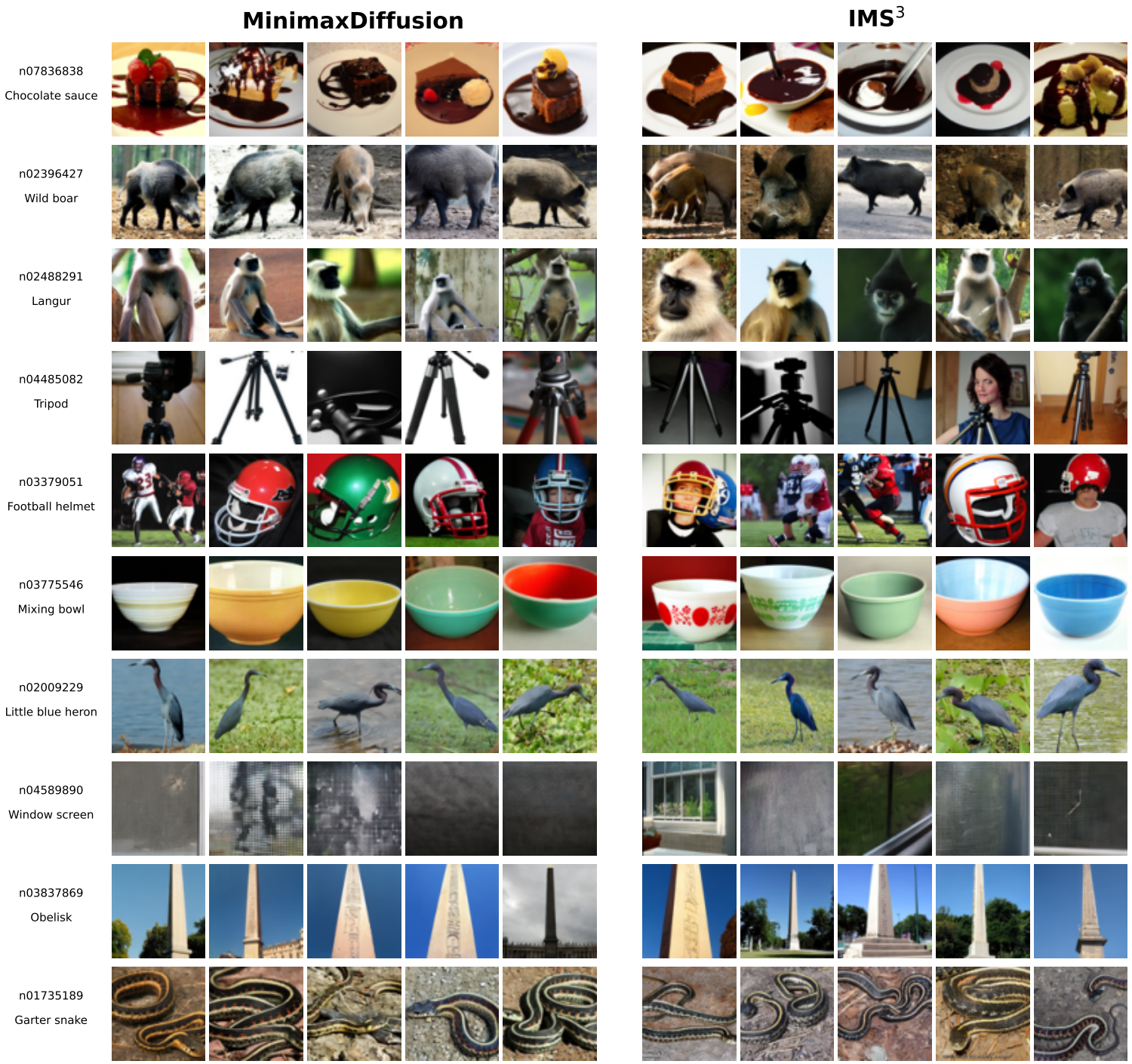


Figure 13. Comparison between samples selected by Minimax [8] (left) and generated by the proposed ImS^3(right) for ImageNet-100 classes 20-29. The class names are marked at the left of each row.

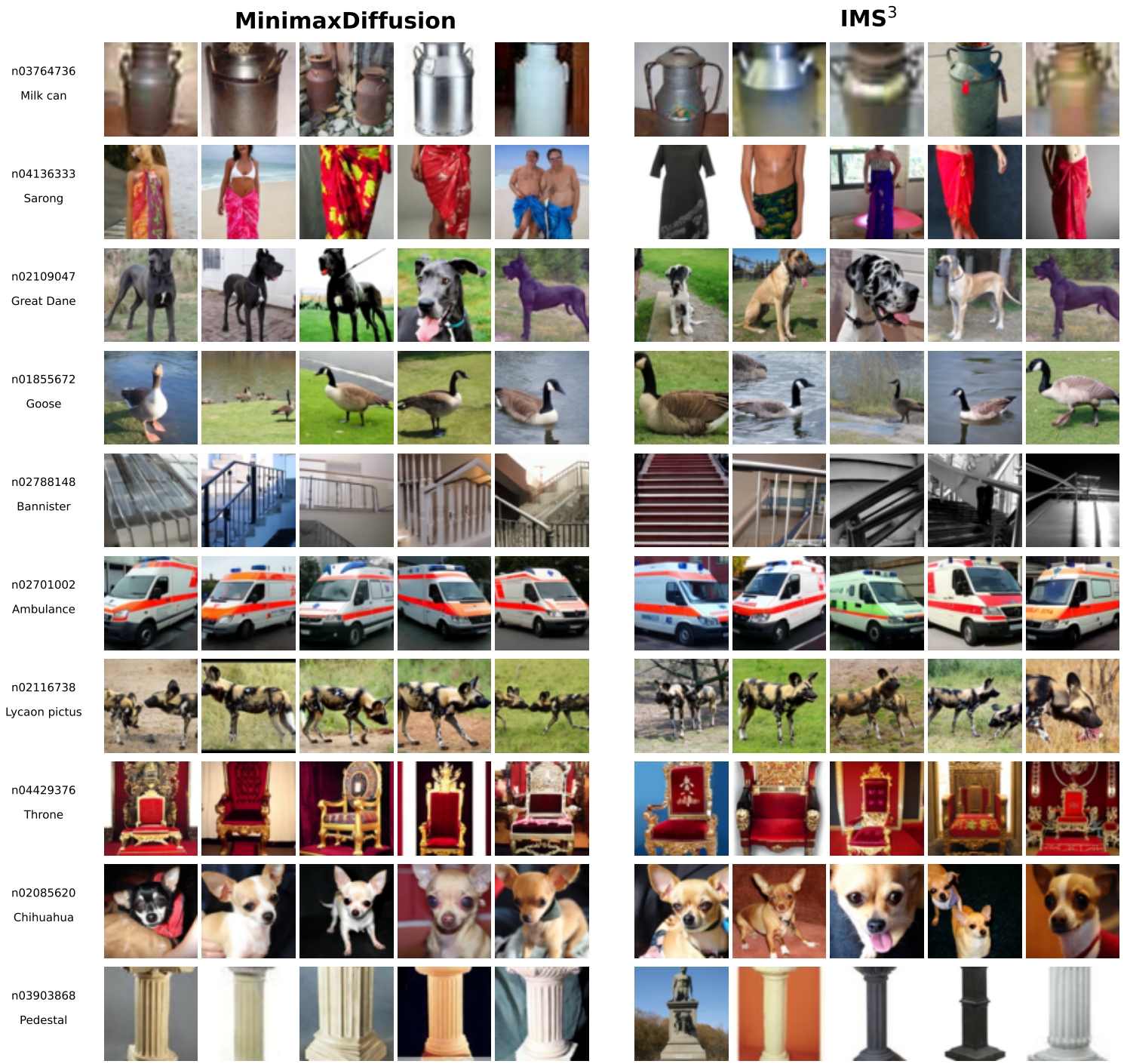


Figure 14. Comparison between samples selected by Minimax [8] (left) and generated by the proposed ImS^3(right) for ImageNet-100 classes 30-39. The class names are marked at the left of each row.

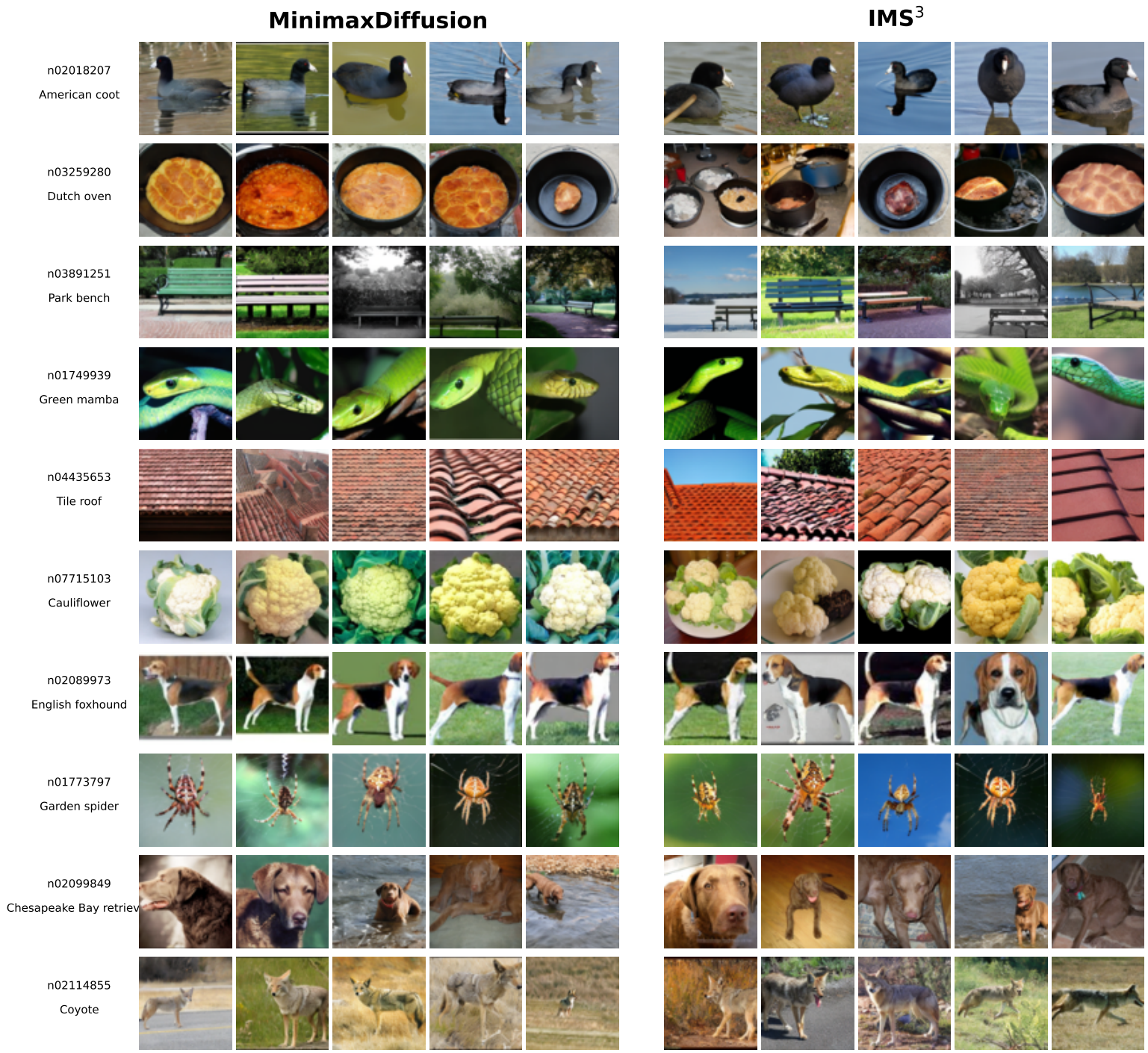


Figure 15. Comparison between samples selected by Minimax [8] (left) and generated by the proposed ImS^3(right) for ImageNet-100 classes 40-49. The class names are marked at the left of each row.

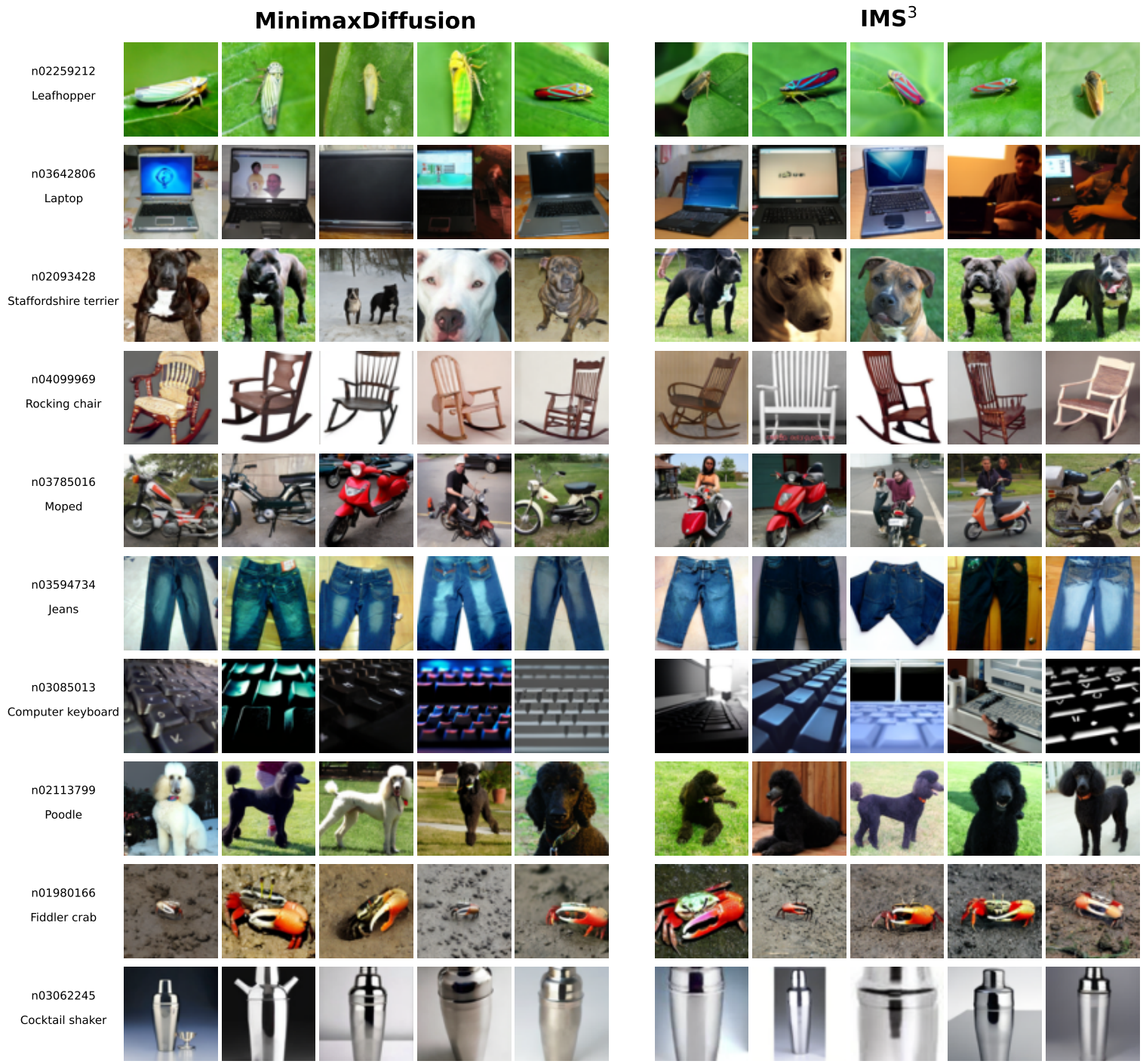


Figure 16. Comparison between samples selected by Minimax [8] (left) and generated by the proposed ImS^3(right) for ImageNet-100 classes 50-59. The class names are marked at the left of each row.

Figure 17. Comparison between samples selected by Minimax [8] (left) and generated by the proposed ImS^3(right) for ImageNet-100 classes 60-69. The class names are marked at the left of each row.

Figure 18. Comparison between samples selected by Minimax [8] (left) and generated by the proposed ImS^3(right) for ImageNet-100 classes 70-79. The class names are marked at the left of each row.

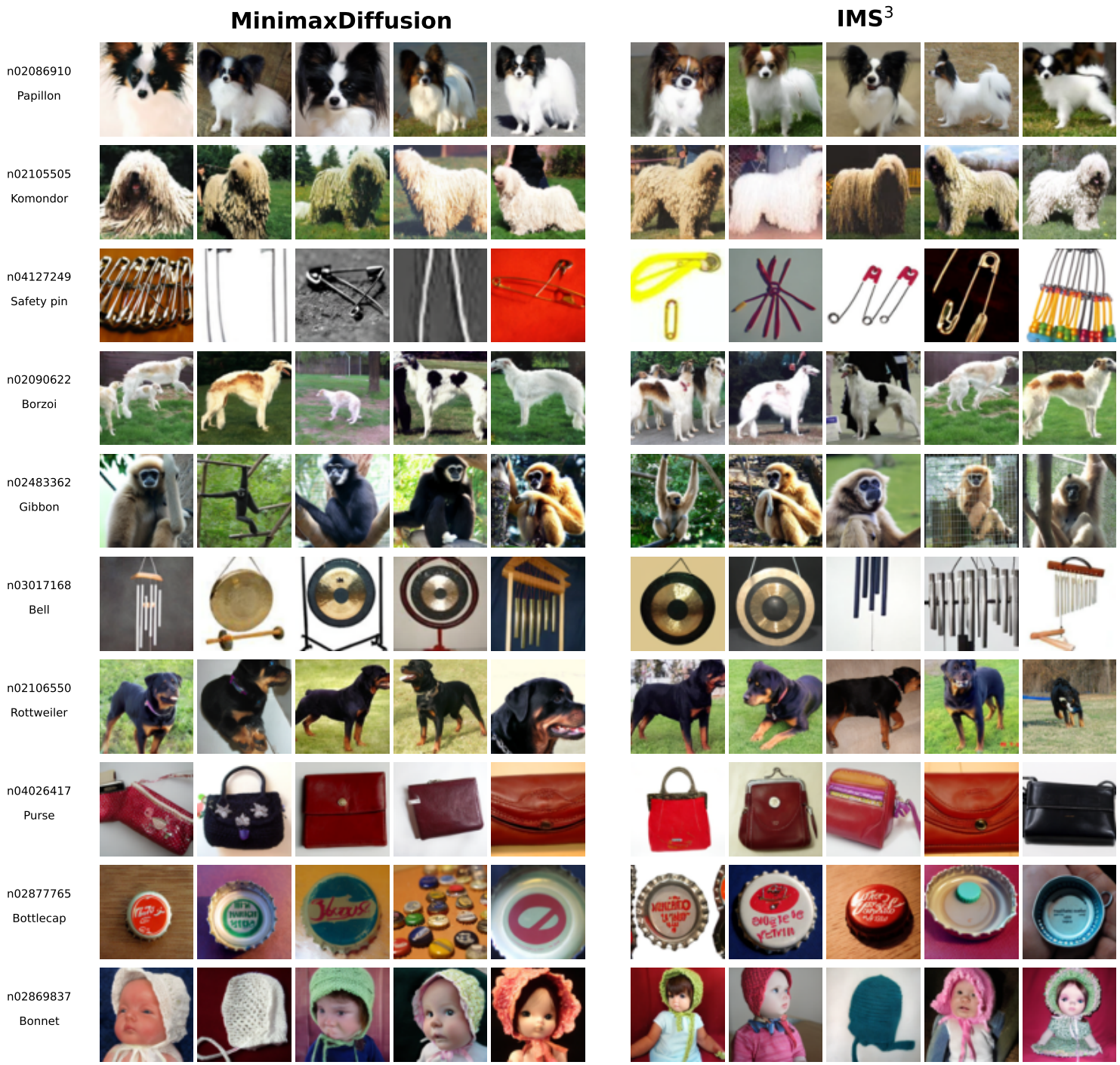


Figure 19. Comparison between samples selected by Minimax [8] (left) and generated by the proposed ImS^3(right) for ImageNet-100 classes 80-89. The class names are marked at the left of each row.

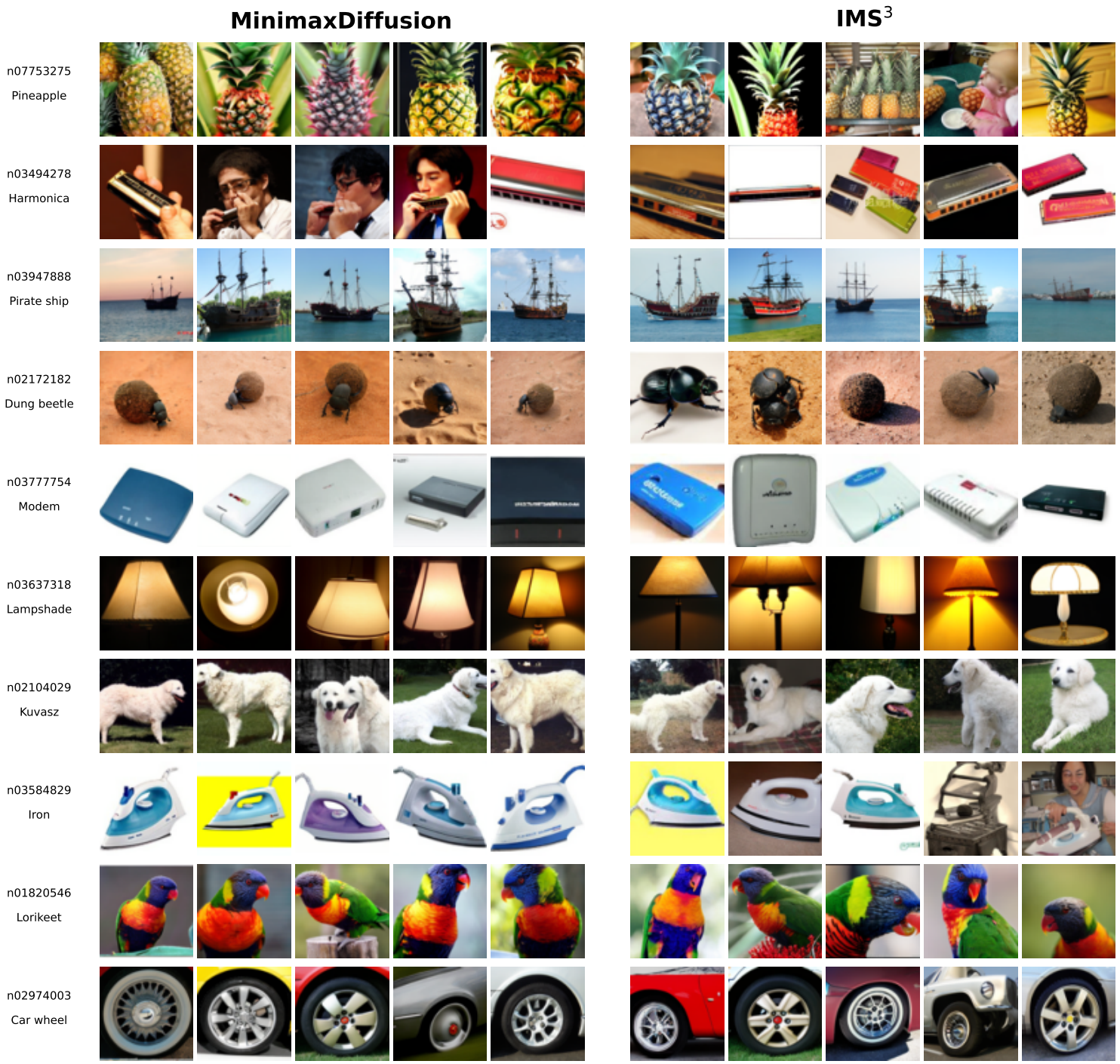


Figure 20. Comparison between samples selected by Minimax [8] (left) and generated by the proposed ImS^3(right) for ImageNet-100 classes 90-99. The class names are marked at the left of each row.