HIERAMP: Coarse-to-Fine Autoregressive Amplification for Generative Dataset Distillation

Supplementary Material

Algorithm 1 Coarse-to-Fine Semantic Amplification

Input: Multi-scale queries $\{Q_n\}_{n=1}^{9}$, keys $\{K_n\}_{n=1}^{9}$, values $\{V_n\}_{n=1}^{9}$; class token is appended as the last query in Q_n, Per-head mask $\{p_n^{(h)}\}$; per-scale key counts $\{L_k^n\}$; heads H, head dim d_h, Stage schedules $(\rho_{1:3}, \rho_{4:6}, \rho_{7:9})$ and $(\beta_{1:3}, \beta_{4:6}, \beta_{7:9})$.
Output: Amplified attentions $\{\tilde{\alpha}_n\}$ and/or reweighted contexts $\{\tilde{O}_n\}$.
for $n = 1$ to 9 **do**
  ▷ coarse (1–3) → mid (4–6) → fine (7–9)
  $(\rho, \beta) \leftarrow \textsc{StageParams}(n)$
  for $h = 1$ to H **do**
    $L_n^{(h)} \leftarrow \dfrac{Q_n^{(h)}(K_n^{(h)})^\top}{\sqrt{d_h}} + p_n^{(h)}$
    ▷ masked logits at scale n
    $\alpha_{n,\text{cls}}^{(h)} \leftarrow \text{Softmax}\big(L_n^{(h)}[-1,\ 1{:}L_k^n]\big)$
    ▷ class → same-scale keys
  end for
  $m_n \leftarrow \frac{1}{H}\sum_{h=1}^{H} \alpha_{n,\text{cls}}^{(h)} \in \mathbb{R}^{1\times L_k^n}$
  ▷ head-avg saliency
  $k \leftarrow \max\big(1, \lfloor \rho \cdot L_k^n \rfloor\big)$
  $S_n \leftarrow \textsc{TopK}(m_n, k)$
  $a_n \in \{0,1\}^{L_k^n}$ initialized to 0
  $(a_n)_j \leftarrow 1$ iff $j \in S_n$
  ▷ binary indicator
  for $h = 1$ to H **do**
    $B_n^{(h)} \leftarrow \beta \cdot \mathbf{1}_{L_q^n+1}\, a_n^\top \in \mathbb{R}^{(L_q^n+1)\times L_k^n}$
    $\tilde{L}_n^{(h)} \leftarrow L_n^{(h)}$
    $\tilde{L}_n^{(h)}[:,\ 1{:}L_k^n] \mathrel{+}= B_n^{(h)}$
    $\tilde{\alpha}_n^{(h)} \leftarrow \text{Softmax}(\tilde{L}_n^{(h)})$
  end for
  $\tilde{O}_n \leftarrow \textsc{AttnOut}\big(\{\tilde{\alpha}_n^{(h)}\}_{h=1}^{H},\ \{V_n^{(h)}\}_{h=1}^{H}\big)$
end for
return $\{\tilde{O}_n\}_{n=1}^{9}$

A. Algorithm

We provide more details about our coarse-to-fine autoregressive amplify algorithm here. As shown in Algorithm 1, we hierarchically amplify the most salient regions at coarse-to-fine scales, yielding semantics that are maximally informative for classification. We then apply the residual rules described in Sec. 3.1 to obtain the final output. In our final configuration, the amplification factor is set to 5 at all scales and is applied to the weights after the softmax.

B. Generalizing to DiT

To demonstrate that HIERAMP is backbone-agnostic, we extended it to DiT [29, 46] and evaluated it on ImageNet-Woof [18]. Specifically, let A denote the attention output at a given transformer block. We apply spatially guided scaling:

$$\tilde{A} = A \odot (1 + \alpha M), \tag{14}$$

where M is the object-region projected to the token space, α controls the amplification strength, and $\odot$ denotes element-wise multiplication. As shown in Table 5, HIERAMP improves the accuracy while maintaining stable generation compared to the vanilla DiT baseline, which indicates that our method generalizes beyond VAR and is compatible with diffusion-based transformer backbones.

Figure 6 shows visual comparisons before and after applying HIERAMP. We observe enhanced object prominence and clearer semantic structures, while background regions remain largely unaffected or much more relevant. The modulation improves object-level consistency without introducing noticeable artifacts, demonstrating the effectiveness of region-aware scaling in diffusion transformers.

Importantly, this extension requires no architectural redesign. Besides, compared to token-level conditioning methods [49–51, 54, 55], our approach operates directly on intermediate feature representations, enabling spatially precise modulation with negligible computational overhead. These findings suggest that HIERAMP provides a general mechanism for hierarchical semantic control across diverse generative backbones.

C. FID Comparisons

Comparison with Prior Methods. We compare the Fréchet Inception Distance (FID) [16] of our method with representative dataset distillation approaches, including Minimax [13] and D^{3}HR [61]. FID is computed against the original ImageNet-1K training set under 10 and 50 images per class (IPC). Lower FID indicates better performance.

As shown in Table 6, our method consistently achieves lower FID scores across both IPC settings. In particular, at 10 IPC, our approach improves FID from 18.3 (Minimax) and 19.0 (D^{3}HR) to 17.3. At 50 IPC, we obtain 13.2, outperforming others. We further analyze whether the hierarchical amplification strategy affects generative fidelity. Table 6 reports FID before and after applying HIERAMP on

IPC	VAR without class tokens	VAR with class tokens
10	45.9 ± 0.3	45.6 ± 0.3
50	59.5 ± 0.1	59.3 ± 0.1

Table 7. **Effect of class tokens on top-1 accuracy on ImageNet-1K.** Models with and without class tokens show comparable performance across different IPC settings.

Model	Ours	D^3HR (DDIM-based, 30 steps)
Time (s/img)	$\mathbf{0.147 \pm 0.001}$	0.456 ± 0.002

Table 8. **Inference latency comparison with DDIM-based method on ImageNet-Woof.**

Model	Base VAR	VAR + cls	VAR + cls + Amp (Ours)
Latency (s/img)	0.139 ± 0.002	0.145 ± 0.002	0.147 ± 0.001
Peak Memory (GB)	1.770 ± 0.000	1.790 ± 0.000	1.840 ± 0.000

Table 9. **Latency and peak memory for VAR-based distillation with incremental modules.**

Dataset	Model	IPC	DiT	DiT + HIERAMP
ImageNet-Woof	ResNet-18	10	41.0 ± 0.6	$\mathbf{43.1 \pm 0.3}$
		50	66.2 ± 0.3	$\mathbf{68.2 \pm 0.3}$

Table 5. **Quantitative comparison of the DiT backbone before and after applying HIERAMP.**

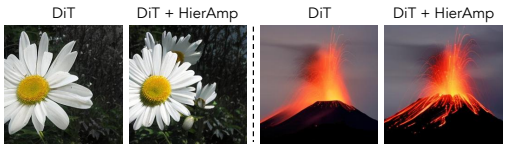

Figure 6. **Qualitative comparison of the DiT backbone before and after applying HIERAMP.**

IPC	Minimax	D^3HR	VAR	Ours
10	18.3 ± 0.2	19.0 ± 0.2	17.5 ± 0.1	$\mathbf{17.3 \pm 0.1}$
50	14.3 ± 0.2	14.9 ± 0.1	$\mathbf{13.1 \pm 0.1}$	13.2 ± 0.1

Table 6. **FID of different dataset distillation methods on ImageNet-1K under 10 and 50 IPC.**

VAR. The results show that FID remains comparable to the default VAR baseline across IPC settings. Specifically, the difference is marginal (e.g., 17.5 vs. 17.3 at 10 IPC).

The above results indicate that HIERAMP preserves visual fidelity while enhancing semantic discriminability. This confirms that the proposed strategy does not degrade generation quality.

D. Effectiveness of Class Tokens

We evaluate the impact of class tokens on both performance and generation quality of VAR. As shown in Tab. 7, models trained with and without class tokens exhibit highly similar top-1 accuracy across different IPC settings, indicating that class tokens introduce no significant advantage or degradation in classification performance.

To further examine their generative behavior, we visualize distilled images produced by both variants in Fig. 7. The "w/o" setting denotes VAR trained without class tokens, while "w/" denotes the standard model with class tokens. Consistent with the quantitative results, the visualizations demonstrate comparable generative capacity: both models produce class-consistent images with similar semantic fidelity and structural detail. These findings show that employing VAR with class tokens is a reliable design choice, and can future provide additional object-focused attention benefits.

E. Computational latency

Comparison with Diffusion Models. We compare the inference speed of our method with a representative diffusion model, DDIM [37] used in D^3HR, under the same setting (batch size = 1). As shown in Table 8, our approach achieves significantly lower latency, processing an image in 0.147 s compared to 0.456 s for DDIM with 30 denoising steps. This efficiency arises from the progressive prediction of scales using fewer tokens and a reduced number of inference steps (≤ 10), demonstrating that hierarchical amplification can accelerate generation without sacrificing quality.

Resource Consumption of Distillation. We further report the computational cost of our dataset distillation pipeline based on VAR. Table 9 shows latency and peak memory for the base VAR, VAR with the class token, and VAR with class token plus hierarchical attention amplification (HIERAMP). The additional overhead introduced by the class token and attention modulation is negligible, with runtime increasing only slightly ($0.139 \rightarrow 0.147$ s/img) and peak memory remaining comparable ($1.770 \rightarrow 1.840$ GB).

These results indicate that HIERAMP provides a computationally efficient alternative to standard diffusion-based generation while maintaining competitive fidelity. The progressive scale prediction and token-efficient design contribute to faster inference, and the pipeline ensures minimal runtime and memory overhead for extended VAR variants.

F. Analysis on More Amplification Combinations

We conducted additional experiments with different attention amplification combinations to further validate the conclusions we draw from Fig. 3: (1) amplify Mid & Fine stages by 5 and Coarse stage by 0.5; (2) amplify Coarse & Fine stages by 5 and Mid stage by 0.5; (3) amplify Coarse & Mid stages by 5 and Fine stage by 0.5. As shown in Fig. 3, amplifying attention at coarse and mid scales increases diversity, with many classes exhibiting higher en-

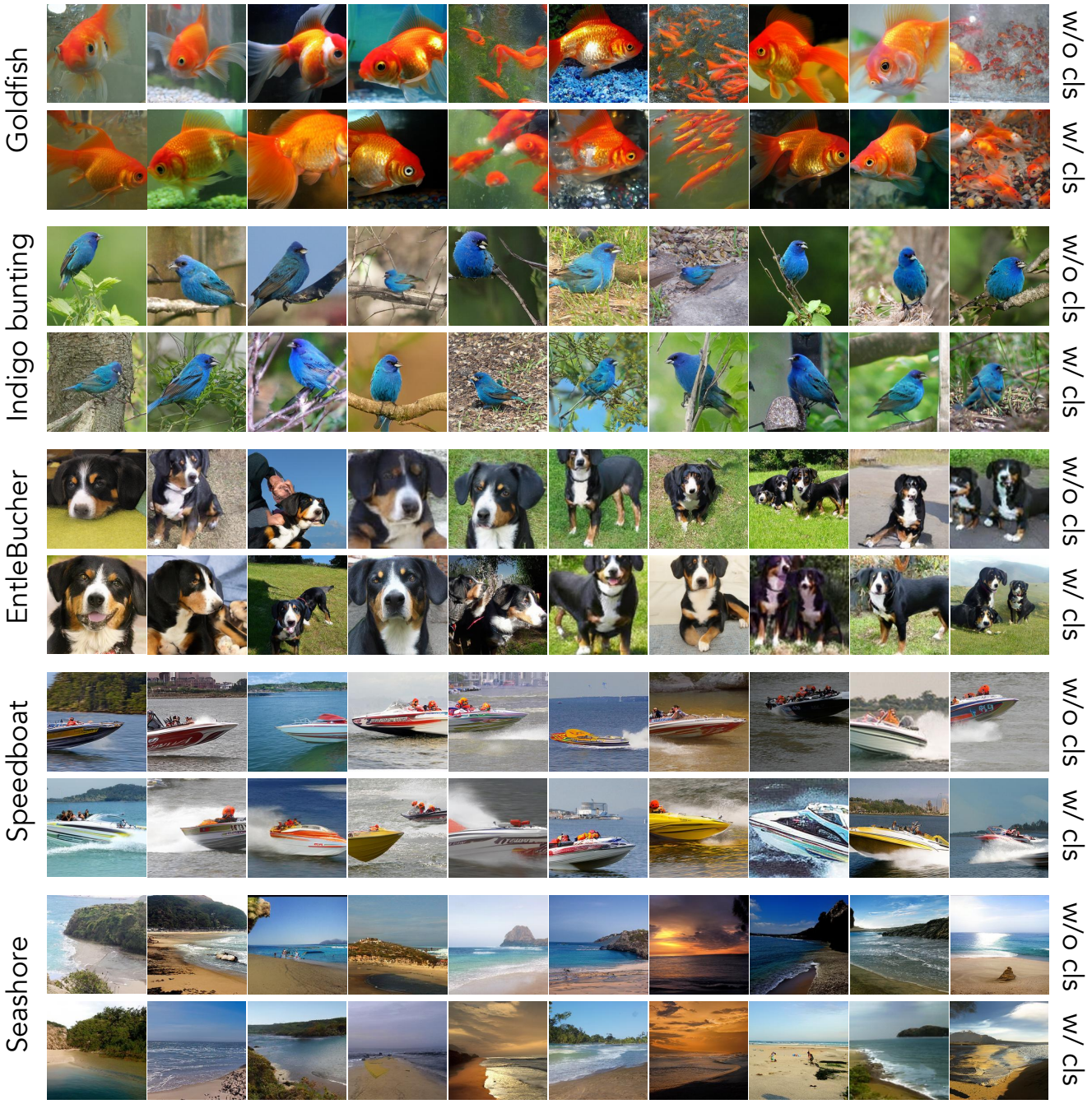

Figure 7. **Visualization of images generated by VAR with and without class tokens on ImageNet-1K, IPC=10.** "w/o" denotes VAR without class tokens, and "w/" denotes VAR with class tokens. The results indicate comparable generative capacity between the two models.

tropy, whereas fine-scale amplification concentrates attention, with many classes exhibiting lower entropy.

For combination (1), which is similar to full-stage amplification (S1–S9) except that the Coarse stage is not amplified by 5, but 0.5, we expect fewer classes to exhibit increased token entropy in S1–S3, consistent with the results in Fig. 8. Similarly, in combination (2), fewer classes show increased token entropy in S4–S6 compared to the full-stage case in Fig. 3. Additionally, reducing the amplification factor in the Fine stage leads to a larger percentage of classes with increased token entropy compared to S7–S9 in Fig. 3,

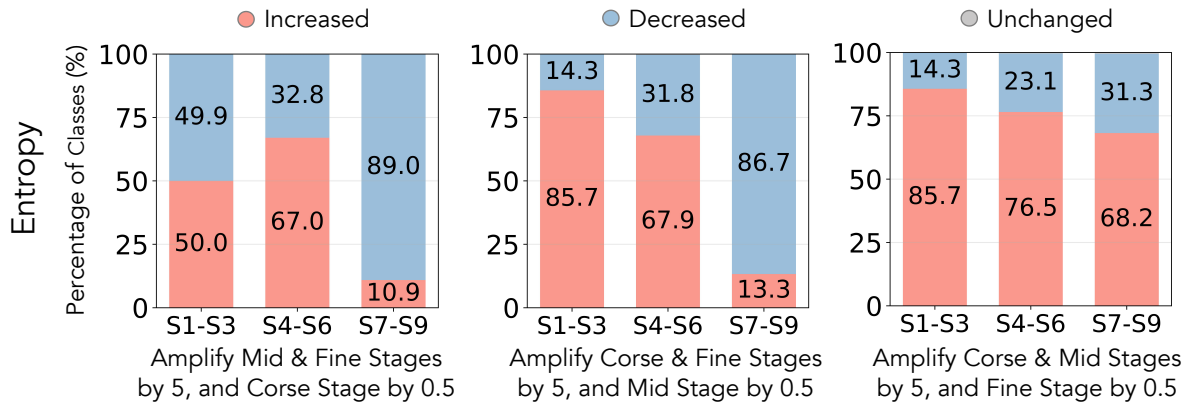

Figure 8. **More amplification combinations impact of attention amplification strategy on token entropy on ImageNet-1K, IPC=50.**

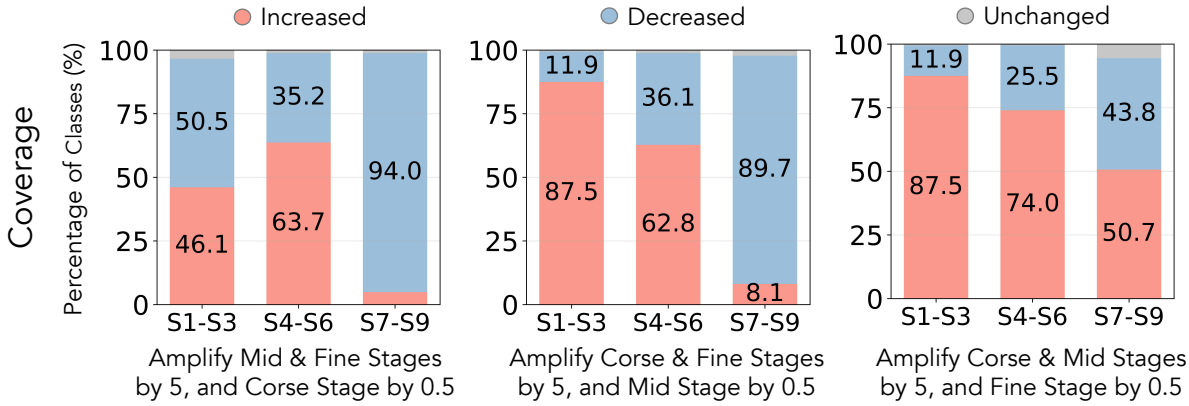

Figure 9. **More amplification combinations impact of attention amplification strategy on token coverage on ImageNet-1K, IPC=50.**

indicating that smaller amplification at fine scales results in less concentrated attention.

The coverage results in Fig. 9 further corroborate these trends. When amplification is applied to the Coarse and Mid stages, coverage expands across a larger portion of the codebook tokens, reflecting greater diversity in the attended regions. In contrast, stronger amplification at the Fine stage leads to more focused and localized attention, reducing overall coverage. For combination (1), the reduced amplification at the Coarse stage results in lower coverage gains in S1–S3 compared to full-stage amplification. Similarly, combination (2) yields smaller coverage increases in S4–S6, aligning with the patterns observed in the entropy analysis. Finally, combination (3), which applies a weaker amplification to the Fine stage, produces broader coverage in S7–S9 relative to the full-stage setting, consistent with the observation that smaller fine-scale amplification reduces attention concentration.

G. Image Visualization and Comparison

We show further visualizations of the distilled images in this section. As illustrated in Fig. 10 and Fig. 11, HIER-AMP generates finer object detail and more diverse objects in a single image, better semantic alignment, and stronger object–background coupling for each class, providing an effective representation of the full dataset.

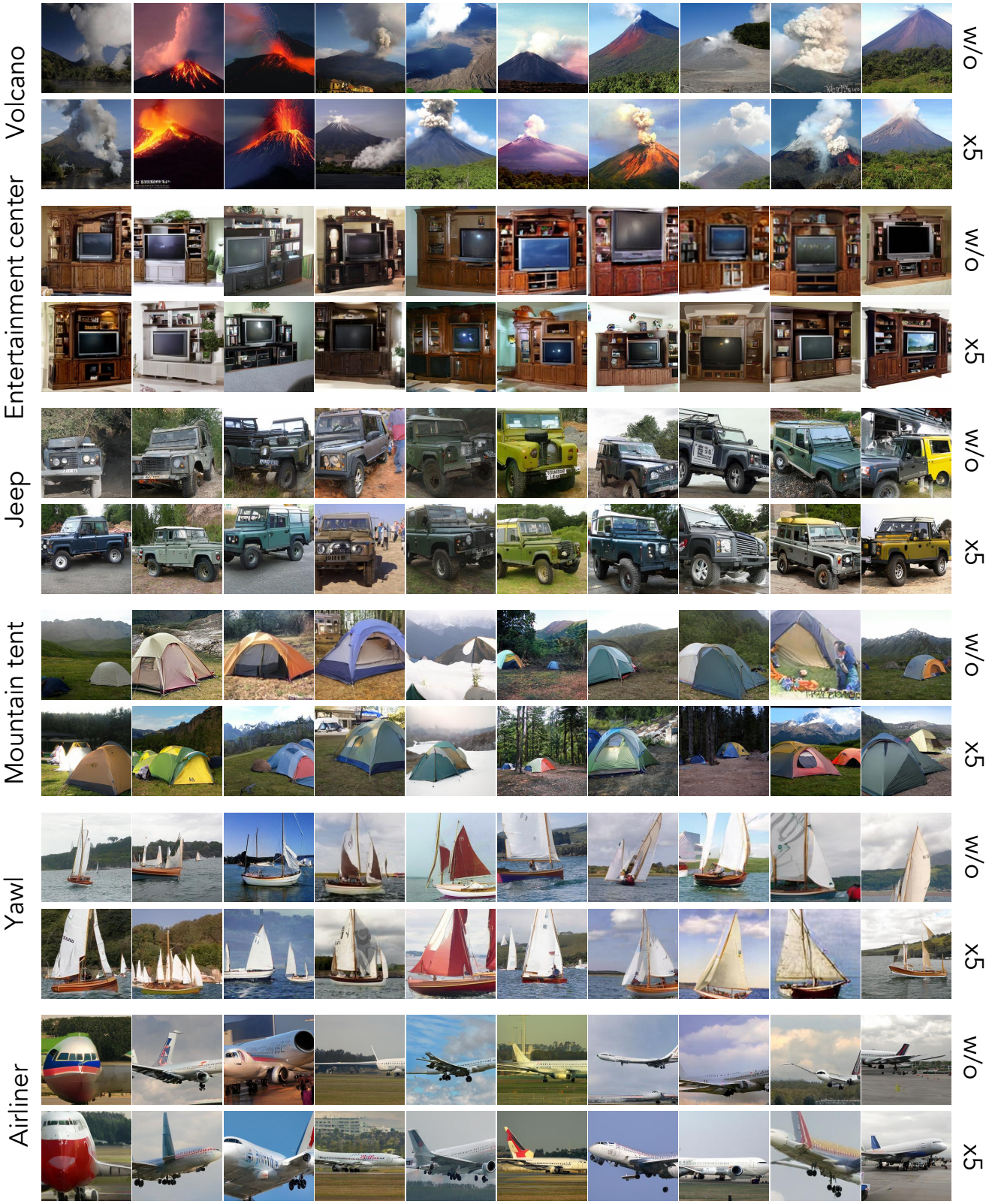

Figure 10. **Visualization of the generated distilled images** (224×224) **on ImageNet-1K, IPC=10.** The first row shows distilled images without amplification (VAR with class tokens). The second row shows distilled images with amplification applied to Full stages (1-9) by a amplification factor of 5.

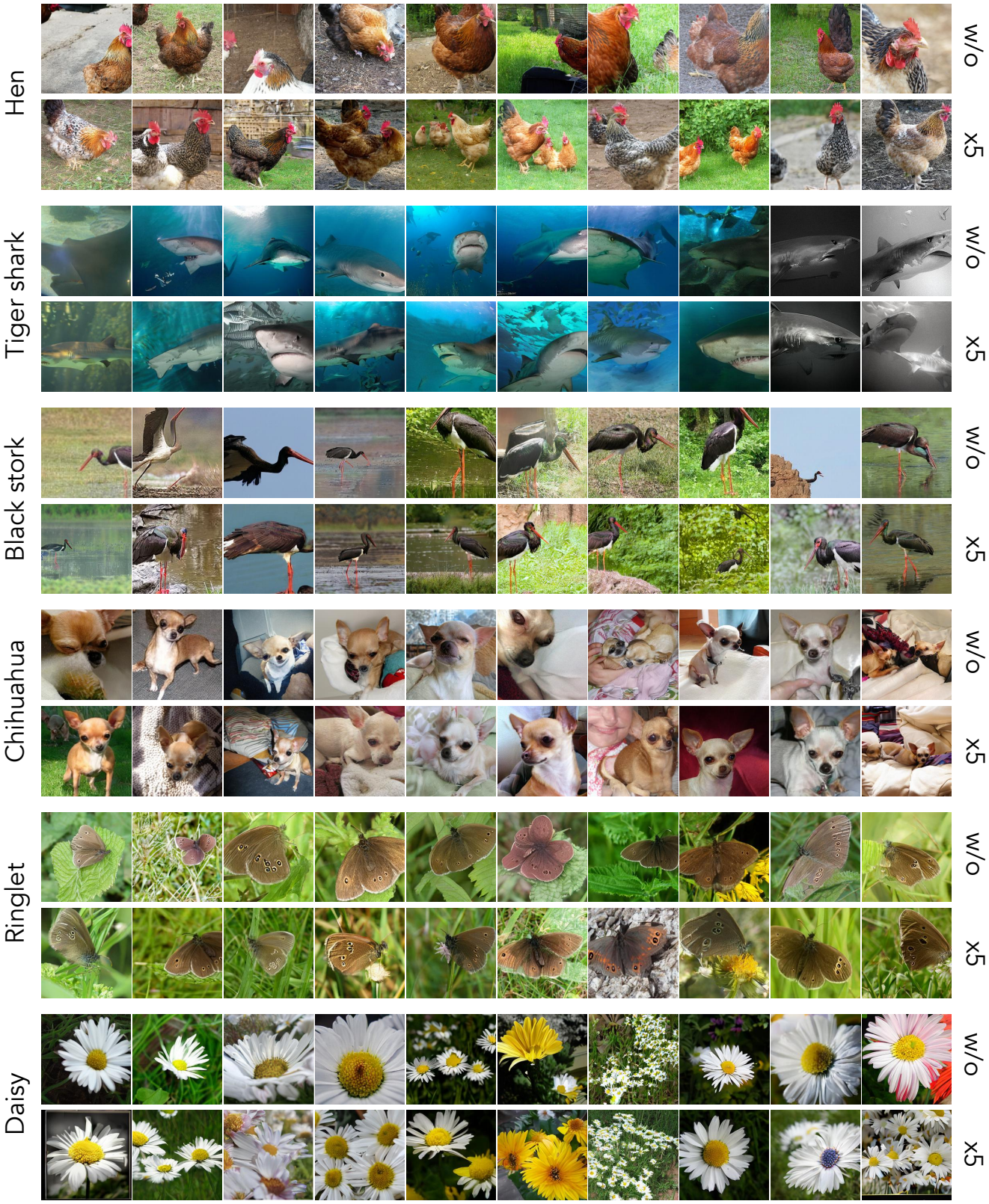

Figure 11. **Visualization of the generated distilled images** (224×224) **on ImageNet-1K, IPC=10.** The first row shows distilled images without amplification (VAR with class tokens). The second row shows distilled images with amplification applied to Full stages (1-9) by a amplification factor of 5.