Gradient-Free Classifier Guidance for Diffusion Model Sampling Supplementary Material

Rahul Shenoy[1*] Zhihong Pan[1*] Kaushik Balakrishnan[1] Qisen Cheng[1]
Yongmoon Jeon[2] Heejune Yang[2] Jaewon Kim[2]
[1]Samsung Display America Lab, USA [2]Samsung Display Co., South Korea

1. Evaluation Metrics

We have explained the reasoning for choosing $\text{FD}_{\text{DINOv2}}$ over FID as the overall image quality metric in the main paper. To further validate this choice, we also conducted a lossy compression test as in [5] to compare FID and $\text{FD}_{\text{DINOv2}}$. As shown in Table 6, For the same 42000 images sampled from SD 1.5, comparing to the original uncompressed output, multiple JPEG compressions are applied with different quality settings. For $\text{FD}_{\text{DINOv2}}$, the metric remains relatively consistent for original ouput as well as different compression. In contrast, as the Bird Species dataset used for assessment consists of JPEG images, FID benefits from applying similar JPEG compression to the original outputs. In fact, FID improves (lower value) continuously with loss of image quality until it reaches the lowest around 80% quality. Lastly, it is shown in Autoguidance [3] that the model and sample settings need to be tuned for optimal FID and $\text{FD}_{\text{DINOv2}}$ separately. As a result, $\text{FD}_{\text{DINOv2}}$ is chosen as the primary sample quality metric for experiments in this work to determine optimal settings.

Table 6. Comparison of different image compression qualities using FID and $\text{FD}_{\text{DINOv2}}$ metrics for 42000 generated samples from SD 1.5 when assessed using the Bird Species dataset.

	Original	JPEG Compression Quality					
	Output	100%	95%	90%	85%	80%	75%
FID↓	13.29	14.20	12.87	10.57	6.48	6.47	6.52
$\text{FD}_{\text{DINOv2}}$ ↓	401.5	400.5	399.2	397.0	397.3	397.6	398.0

2. Additional Implementation Details

2.1. Class-Conditional Generation: Pseudo Code

For our class-conditional experiments, we used the 2[nd] order deterministic sampler from EDM (i.e., Algorithm 1 in [2]) in all experiments with $\boldsymbol{\sigma}(\mathbf{t}) = \mathbf{t}$ and $s(\mathbf{t}) = 1$. We used the default settings $\boldsymbol{\sigma}_{min} = 0.002$, $\boldsymbol{\sigma}_{max} = 80$, $\rho = 7$ and $N = 32$. Note that we use $\boldsymbol{\sigma}$ and $\mathbf{t}$ for EDM to avoid confusion as σ and t are also used in our formulations. To follow the terminology in Algorithm 1, N in original EDM is denoted as T, sampling steps $i = 0, 1, \ldots$ is denoted as $t = T, T-1, \ldots$ and noise schedule $\boldsymbol{\sigma}(\boldsymbol{t}_0), \boldsymbol{\sigma}(\boldsymbol{t}_1), \ldots$ is denoted as $\sigma_T, \sigma_{T-1}, \ldots$ and the noise schedule is calculated as follows:

$$\sigma_t = \begin{cases} \left(\boldsymbol{\sigma}_{max}{}^{\frac{1}{\rho}} + \frac{T-t}{T-1}(\boldsymbol{\sigma}_{min}{}^{\frac{1}{\rho}} - \boldsymbol{\sigma}_{max}{}^{\frac{1}{\rho}})\right)^{\rho} & t > 0 \\ 0 & t = 0 \end{cases} \tag{8}$$

Equation 4 for $\widehat{x}_0$ estimation is replaced by a multi-step denoising process, as also discussed in Section 4.1. This modification is detailed in Algorithm 2. Although this introduces a few additional NFEs, the impact is minimal since the multi-step estimation is only required once. For instance, in the case of GFCG_{ATG}, the parameters in Algorithm 2 are set as $M_B = \text{ATG}$, $\boldsymbol{\sigma}'_{min} = 0.002$, $\rho' = 7$ and $N' = 4$. As for s_{cp}, it is set to the maximum so that only one classifier prediction is used for all mixed and additive GFCG methods included in Table 2.

2.2. Text-to-Image Generation: Text Prompts

For the main quantitative experiments of text-to-image generations using GFCG and other gradient-free guidance methods, a set of generic prompts are used based on the realistic distribution of the Birds Species dataset. This is designed to minimize the bias between the real and generated images so the $\text{FD}_{\text{DINOv2}}$ metric could be more reliable in quantitative assessment. Each of the Set of following 8 generic text prompts was used to generate 10 samples for each bird species and quantitative results are reported in Table 4.

- *a close up photo of a bird, [bird species]*
- *a close up bird photo, [bird species]*
- *a close up picture of a bird, [bird species]*
- *a close up bird picture, [bird species]*
- *a full body photo of a bird, [bird species]*
- *a full body bird photo, [bird species]*
- *a full body picture of a bird, [bird species]*
- *a full body bird picture, [bird species]*

*Both authors contributed equally to this work.

Algorithm 2: Gradient-free Classifier Guidance for EDM2 with Multi-step Denoising for $\widehat{x}_0$

Input: (i) Trained diffusion models D_θ^m and D_ϕ^g;
(ii) Base Sampling Method $M_B \in \{\text{NG, CFG, SEG, ATG}\}$;
(iii) Noise schedule σ_t calculated using Equation 8 with parameters $(\boldsymbol{\sigma}_{min}, \boldsymbol{\sigma}_{max}, \rho, T)$;
(iv) Parameters for $\widehat{x}_0$ estimation using multi-step denoising $(\boldsymbol{\sigma}'_{min}, \rho', T')$;
(v) Hyperparameters α, β, t_s & s_{cp};
(vi) Trained classifier, $\mathcal{C}$.
Output: A generated (noise-free) image, x_0.

$x_T \sim \mathcal{N}(0, \sigma_T^2 \mathbf{I})$
Desired class: $c_{des} \sim (c_1, c_2, \ldots, c_N)$
for $t = T, T-1, \ldots, 1$ **do**
  $use_{\text{BASE}} \leftarrow$ TRUE;
  if $t \leq t_s$ **then**
    if $(t_s - t)\ \%\ s_{cp} == 0$ **then**
      $\boldsymbol{\sigma}'_{max} \leftarrow \sigma_t$;
      Compute noise schedule $\sigma'_{t'}$ with parameters $(\boldsymbol{\sigma}'_{min}, \boldsymbol{\sigma}'_{max}, \rho', T')$ using Equation 8;
      $\tilde{x}_{T'} \leftarrow x_t$;
      for $t' = T', \ldots, 1$ **do**
        Evaluate $\tilde{x}_{t'-1}$ using Heun's solver and $\widehat{D}'$ computed based on M_B (refer Algorithm 1 in [2]);
      end
      $\widehat{x}_0 \leftarrow \tilde{x}_0$ (consistent with Algorithm 1 in main paper);
      Use $\mathcal{C}$ to estimate $p(c|\widehat{x}_0)$;
      Evaluate ω using Equation 5;
      Estimate the reference class, c_{ref};
    $D_1 \leftarrow D_\theta^m(x_t, \sigma_t, c_{des})$;
    $D_2 \leftarrow D_\phi^g(x_t, \sigma_t, c_{ref})$;
    $\widehat{D} \leftarrow \omega D_1 - (\omega - 1) D_2$;
    $use_{\text{BASE}} \leftarrow$ FALSE;
  if use_{BASE} **then**
    Compute $\widehat{D}$ based on M_B;
    Example: if $M_B =$ ATG then
      $D_1 \leftarrow D_\theta^m(x_t, \sigma_t, c_{des})$, $D_2 \leftarrow D_\phi^g(x_t, \sigma_t, c_{des})$ and $\widehat{D} \leftarrow \omega_{ATG} D_1 - (\omega_{ATG} - 1) D_2$;
  Evaluate x_{t-1} using Heun's solver and $\widehat{D}$ (refer Algorithm 1 in [2]);
end
return x_0

A set of detailed text prompts was used to generate samples, 5 per prompt for each species, for the ablation study reported in Table 5 of the main paper. These detailed prompts are not realistic for all species, as the roadrunner in Figure 1 doesn't perch on tree branches in real life. Besides, the descriptions below only cover part of all natural habitats. As $\text{FD}_{\text{DINOv2}}$ is not applicable for this test given these biases between two distributions, only Precision scores are reported.

- *a photo of a bird perching on a tree branch with flowers blooming around it, [bird species]*
- *a close up photo of a flying bird with fish in its claws, [bird species]*
- *a photo of a bird eating red berry when standing on a rock, [bird species]*
- *a photo of a bird walking on the beach on a raining day, [bird species]*
- *a photo of a bird, [bird species], perching on a tree branch with flowers blooming around it*
- *a close up photo of a flying bird, [bird species], with fish in its claws*
- *a photo of a bird, [bird species], eating red berry when standing on a rock*
- *a photo of a bird, [bird species], walking on the beach on a raining day*

2.3. SEG implementation details

For the implementation of SEG in the EDM2 codebase, we consider σ = 100 for the Gaussian Blur. This is applied in the EDM2 UNet's following blocks in the guidance network:

```
8x8_block1
8x8_block2
8x8_in0
8x8_block0
```

Other values of σ (=10 and 1e6) were also considered but the observations are very similar.

The SEG implementation in SD is similar to [1].

3. More Experimental Results

3.1. Effects of Random c_{ref}

Unlike ATG, which uses the target class as its reference, GFCG leverages a pretrained classifier to adaptively select c_{ref} as the most confusing alternative class, while adjusting ω based on classifier confidence. By contrast, CFG requires training an additional unconditional model and fixes c_{ref} as the empty class. An obvious baseline is to choose c_{ref} randomly at each time step. Results in Table 7 show that, although the $\text{FD}_{\text{DINOv2}}$ is comparable to GFCG, precision drops significantly. This is expected, since random selection provides no consistent guidance direction, leading to reduced fidelity.

Table 7. Study impact of setting random c_{ref} per time-step using $\text{FD}_{\text{DINOv2}}$ and Precision metrics for 50000 generated samples from EDM2-S, assessed with the ImageNet dataset. The best in each metric is highlighted in **bold**

ω =	1.6	2.0	2.45	2.8	3.2
$\text{FD}_{\text{DINOv2}}\downarrow$	61.54	47.22	40.99	**39.89**	41.28
Precision↑	85.8%	87.2%	87.8%	**88.1%**	88.0 %

3.2. Effects of Random Seed Variation

The results presented in Table 1 and 2 of the main paper were generated using the same random seed for image generation. As the random seeds used in the ATG study [3] were not disclosed, we were unable to exactly replicate the reported $\text{FD}_{\text{DINOv2}}$ metric. To demonstrate that the improvement in the $\text{FD}_{\text{DINOv2}}$ metric is independent of random seed choice, we have illustrated the variation in $\text{FD}_{\text{DINOv2}}$ and Precision with random seeds for both ATG and GFCG_{NG} methods in Figure 6. The results show a clear distinction between the two methods in terms of Precision and $\text{FD}_{\text{DINOv2}}$ metrics, indicating that GFCG_{NG} consistently outperforms ATG, regardless of the random seed used.

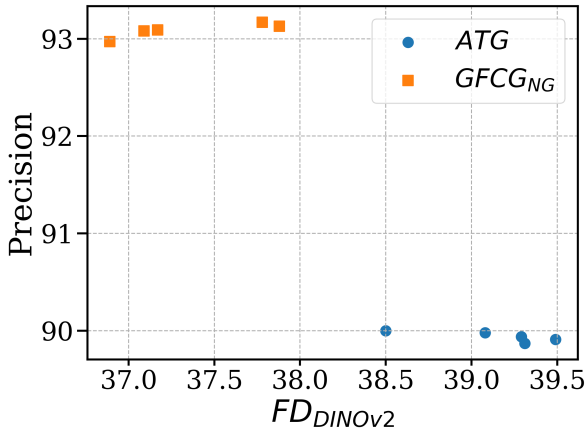


Figure 6. Ablation study for impact on Precision and $\text{FD}_{\text{DINOv2}}$ metric against random seed variation.

3.3. Effects of Guidance Model

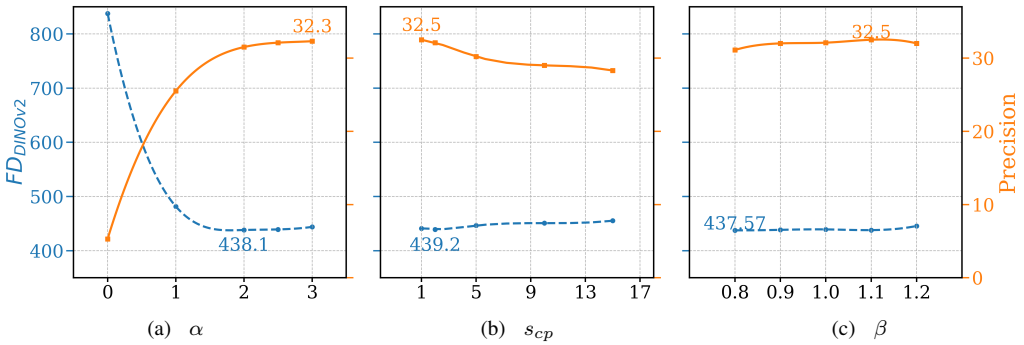
The experiments detailed in the main paper involving the EDM2-S and EDM2-XXL models utilize guidance models (XS, T/16) and (M, T/3.5) respectively for the GFCG and SEG experiments, similar to ATG [3]. These guidance models, with reduced capacity and training, are readily accessible thanks to the publicly available EDM2 codebase [4]. However, this availability may not extend to other class-conditional or text-to-image generation diffusion models. Table 9 presents the $\text{FD}_{\text{DINOv2}}$ and Precision metrics for the GFCG_{NG} method, based on the capacity and training of the guidance model. The hyperparameters for GFCG, α, β, and t_s, are set to 0.85, 1.25, and 17, respectively. Similar to ATG, reducing training significantly impacts the $\text{FD}_{\text{DINOv2}}$ metric, while reducing capacity only results in the worst performance. The highest precision is achieved when using the same guidance model as the main model with some degradation in $\text{FD}_{\text{DINOv2}}$ metric.

Table 8. Study impact of guidance model capacity and training for GFCG_{NG} method using $\text{FD}_{\text{DINOv2}}$ and Precision metrics for 50000 generated samples from EDM2-S, assessed with the ImageNet dataset. The best in each metric is highlighted in **bold** and the second best is marked with underline.

	$\text{FD}_{\text{DINOv2}}\downarrow$	Precision↑	M_g	EMA_m	EMA_g
Reduce capacity	51.21	93.3%	(XS,T)	0.085	0.085
Reduce training	<u>39.36</u>	<u>93.6%</u>	(S,T/16)	0.085	0.170
Same both	47.79	**94.0%**	(S,T)	0.085	0.085
Reduce both	**36.89**	93.0%	(XS,T/16)	0.085	0.165

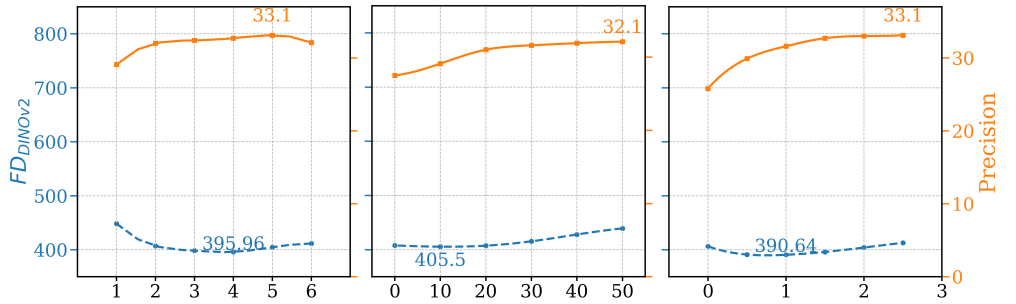
3.4. Effects of Classifier Model

Different classifier models, with varying sizes and top-1 and top-5 accuracies on ImageNet-1k (acc@1 and acc@5 in Table 9), were considered in the ablation study for classifier predictions in GFCG-based methods. ResNet-18, which is one-fourth the size of ResNet-101 used in the



(a) α (b) s_{cp} (c) β

Figure 7. Ablation studies for GFCG method: text-to-image generations (8,400 samples)



(a) ω_{CFG} (GFCG+CFG) (b) t_s (GFCG_{CFG}) (c) α (GFCG+CFG)

Figure 8. Additional ablation studies for text-to-image generations (8,400 samples)

main tests, achieved a high precision of 92.8% compared to ATG's 90.0%, while maintaining a similar $\text{FD}_{\text{DINOv2}}$ to ATG. ResNet-101 exhibited comparable $\text{FD}_{\text{DINOv2}}$ and precision metrics to ResNet-152, as shown in Table 9, but with a smaller model size, and was thus selected for the main experiments in the paper.

3.5. Effects of $\widehat{x}_0$ Estimation Methods

For all the mixed and additive GFCG methods presented in Table 2, s_{cp} is set to its maximum value and T' is set to 4 for $\widehat{x}_0$ estimation. Although this introduces 7 additional NFEs, which is minimal compared to the 63 NFEs, it results in a significant boost in precision and some improvement in $\text{FD}_{\text{DINOv2}}$ as well (see Table 2). We explore two methods to further reduce the NFEs. The first method is to reduce the number of steps for $\widehat{x}_0$ estimation. The second method is to use a smaller and lower-trained guidance model as the main model for $\widehat{x}_0$ estimation. For instance, the guidance model used in the majority of the EDM2-S experiments is EDM2-XS, trained for T/16. If M_B = ATG, then line 32 in Algorithm 2 would change to $D_1 \leftarrow D^g_\phi(x_t, \sigma_t, cdes)$, $D_2 \leftarrow D^g_\phi(x_t, \sigma_t, cdes)$ and $\widehat{D} \leftarrow \omega_{ATG} D_1 - (\omega_{ATG} - 1)D_2$, which essentially applies the NG method using the guidance model only for $\widehat{x}_0$ estimation. As a smaller model is used for $\widehat{x}_0$ estimation, we ignore the NFEs added by this method. The results of these two methods compared to the main paper results are presented in Table 10.

Table 9. Study impact of choice of classifier model for GFCG_{NG} method using $\text{FD}_{\text{DINOv2}}$ and Precision metrics for 50000 generated samples from EDM2-S, assessed with the ImageNet dataset. The best in each metric is highlighted in **bold** and the second best is marked with underline.

Classifier	$\text{FD}_{\text{DINOv2}}\downarrow$	Precision↑	Mparams	Gflops	acc@1	acc@5
ResNet-18	37.32	92.8%	11.7	1.81	69.8%	89.1%
ResNet-34	37.02	92.8%	21.8	3.66	73.3%	91.4%
ResNet-50	38.03	92.9%	25.6	4.09	80.9%	95.4%
ResNet-101	<u>36.89</u>	<u>93.0%</u>	44.5	7.80	81.9%	95.8%
ResNet-152	**36.74**	**93.1%**	60.2	11.51	82.3%	96.0%

Table 10. Study impact of methods to reduce NFEs for $\widehat{x}_0$ estimation for GFCG_{NG} method using $\text{FD}_{\text{DINOv2}}$ and Precision metrics for 50000 generated samples from EDM2-S, assessed with the ImageNet dataset. The best in each metric is highlighted in **bold** and the second best is marked with underline.

	$\text{FD}_{\text{DINOv2}}\downarrow$	Precision↑	σ'_{min}	T'	NFEs
Method 1	40.36	92.5%	-	1	64
	37.32	<u>92.9%</u>	1.0	2	66
Method 2	**36.35**	92.3%	0.002	4	63
Main Paper	<u>36.89</u>	**93.0%**	0.002	4	70

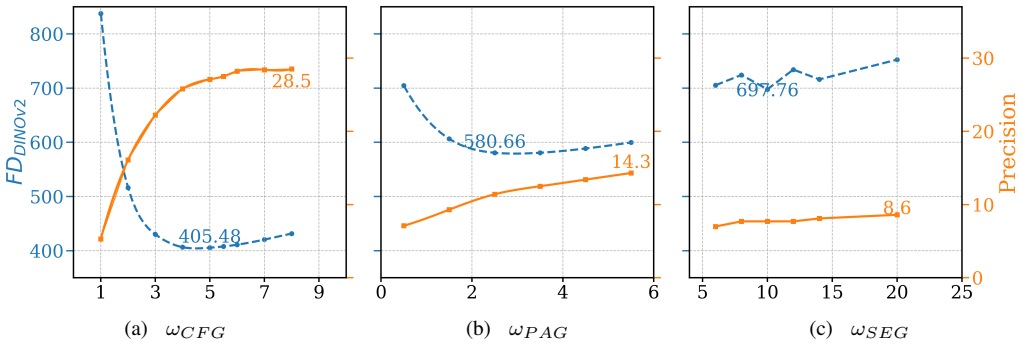


Figure 9. Ablation studies for other guidance methods reported in Table 4 for text-to-image generations

3.6. Text-to-Image: Stochastic Reference Class Sampling

As explained in Equation 7, a stochastic reference class can be sampled each time a classifier prediction is applied. It improves sample quality when there are frequent classifier predictions, i.e. s_{cp} is small. Based on that, the experimental results of text-to-image generations in Table 4 are conducted with this enabled. For comparison, we compare GFCG methods with stochastic reference class sampling to their counterparts with determinist reference class and the quantitative results are included in Table 11. It shows that the stochastic methods are better than their deterministic counterparts in overall performance considering both $\text{FD}_{\text{DINOv2}}$ and Precision.

Table 11. Study impact of stochastic reference class sampling comparing to deterministic reference class in text-to-image generations.

Method	c_{ref}	$\text{FD}_{\text{DINOv2}}\downarrow$	Precision↑
GFCG	Stochastic	418.8	32.3%
	Deterministic	428.8	32.4%
GFCG_{CFG}	Stochastic	392.3	30.2%
	Deterministic	394.7	29.1%
GFCG+CFG	Stochastic	379.2	32.4%
	Deterministic	377.6	31.6%

3.7. Text-to-Image: CLIP-based Guidance

A pre-trained bird species classifier was used for the text-to-image generation experiments in the main paper. While it improves the performance in classification accuracy significantly, it limits it's general application as a pretrained classifier is needed. An experiment was also conducted to use an existing CLIP model, ViT-L/14-DFN [1] as a zero-shot classifier for guidance. As shown in Table 12, while it improves Precision score comparing to CFG, it comes with a significant trade-off in $\text{FD}_{\text{DINOv2}}$. Comparing to GFCG using a pre-trained classifier, it lags behind in both metrics. This is understandable as generic CLIP is not well suited for fine-grained classification like 525 different bird species.

[1] https://github.com/mlfoundations/open_clip

Table 12. Comparison of CLIP-based GFCG and other methods using 42,000 generated samples from SD 1.5, assessed with the Bird Species dataset.

	$\text{FD}_{\text{DINOv2}}\downarrow$	Precision↑	ω_{CFG}	ω_{PAG}	ω_{SEG}	α	t_s
CFG	394.0	27.3%	5.5	-	-	-	-
GFCG	418.8	32.3%	-	-	-	2.5	50
GFCG-CLIP	461.6	28.9%	-	-	-	3.5	50

3.8. Text-to-Image: Ablation Studies

A set of ablation studies were conducted to determine the optimal settings for GFCG based sampling and three are shown in Figure 7. For GFCG, increasing α improves Precision consistently but $\text{FD}_{\text{DINOv2}}$ starts to worsen when it is beyond 2. As evident, α has the highest impact and β the least. For s_{cp}, the Precision value is the highest when it is set as the minimum of 1, while achieving the lowest $\text{FD}_{\text{DINOv2}}$ too.

For the mixed GFCG_{CFG} method, the only key variable is t_s as ω_{CFG} for CFG and α for GFCG are using the same optimal setting of each respectively. For the additive method of GFCG+CFG, ω_{CFG} and α are investigated to find the optimal settings. As shown in Figure 8, the optimal values are around 4.0 and 1.5, lower than the settings of 5.5 and 2.0 when optimized for CFG and GFCG individually.

For fair comparison, we also study the effect of the guidance scale ω for CFG, PAG and SEG in Figure 9. For CFG, ω has a significant role in terms of the $\text{FD}_{\text{DINOv2}}$ and Precision of the generated images. For PAG, ω has a lesser influence, and for SEG the impact is the least. Among these three guidance methods, CFG is much superior than PAG and SEG in terms of $\text{FD}_{\text{DINOv2}}$ and Precision.

4. Class-Conditional Visual Examples

Visual examples from class-conditional image generation using existing guidance methods (refer to Table 1) are compared with our GFCG method in EDM2-S sampling, as illustrated in Figure 10. Additionally, we compare GFCG to

the additive method GFCG$_{\text{ATG}}$+CFG, which achieves state-of-the-art performance in FD$_{\text{DINOv2}}$ for EDM2-S. Further visual examples can be found in Figure 17. The visual results corroborate the quantitative metrics for Precision, with GFCG-generated images demonstrating a strong alignment with class labels in most cases. This is evident in the examples of *mushroom* and *collie* in Figure 10, where other guidance methods often confuse *mushroom* with *agaric* and *collie* with *border collie*. However, this precise alignment does result in a slight trade-off in diversity, as seen in the *orange* class example, where GFCG generates a zoomed-in image of a single orange. The additive method GFCG$_{\text{ATG}}$+CFG mitigates this issue by balancing diversity and class accuracy, as illustrated in the *orange* class example in Figure 10, where it generates a bunch of oranges in a basket.

To further explore the differences in diversity between GFCG and mixed methods, we compare GFCG with GFCG$_{\text{ATG}}$, which achieves state-of-the-art FD$_{\text{DINOv2}}$ for EDM2-XXL. We present visual examples displaying 10 images per class for selected classes in Figure 11, with additional examples in Figure 18. Figure 11 shows that while GFCG-generated samples align closely with class labels (notably in the *collie* and *orange* cases, which other methods confuse with *border collie* and *lemon*), there is a modest reduction in diversity. For instance, in the *orange* class, GFCG tends to zoom in on the oranges or exclude other fruits, compared to the mixed method. GFCG may also remove or modify background objects which may cause confusion with the target class, as seen in the *pizza* and *valley* classes in Figure 11. Mixed methods, particularly with ATG, help preserve diversity while enhancing class accuracy.

5. Text-to-Image Visual Examples

For text-to-image generations, the results presented in the main paper were all based on samples from SD 1.5 and more visual examples are included here. Additionally, we also conducted experiments using another popular model, DeepFloyd IF model[2] from Stability AI. Some examples are included in Section 5.3.

5.1. Generic Text Prompts

Some visual examples from text-to-images generation using the set of generic prompts are included in Figure 12. The probability of classifying each generated sample as the target class is also included for reference. In general, the GFCG results have the best class accuracy. The gained accuracy could be caused by compositional change in the first example, as well as correct anatomic features like feather color in the second. For GFCG$_{\text{CFG}}$, it often maintains the overall composition of CFG and is possible to improve class accuracy even when the changes are negligible with untrained eyes like the first example. In the case of the last example, minor changes like chest patterns in GFCG$_{\text{CFG}}$ result in significantly increased probability too.

As shown in Figure 13, GFCG results may end up worse than other methods too. For the first example of *TIT MOUSE*, the text-to-image model obviously has misunderstood *MOUSE* without recognizing its context as a bird specie. In the case of CFG, as it is guided away from an unconditional model, it will enhance the wrong features associated with MOUSE, as well as correct features associate with other keywords like *bird*. For GFCG, as the enhancement is in the reference to a photo of another bird species, the difference will be focused between *TIT MOUSE* and another bird species which results in more prominent mouse features. Similar failure happens in the second example where color features related to *TEAL* is magnified.

More visual examples using generic text prompts are shown in Figure 19 at the end.

5.2. Detailed Text Prompts

Some visual examples using the set of detailed prompts are included in Figure 14. Note that the hyperparameters like α and t_s for GFCG related methods were optimized for the generic prompts and adopted for detailed prompts without further tuning. It shows that GFCG has the best class accuracy while preserving the overall accuracy of the full text prompt, including improving large features like the first example, or small details like the last one. It is noted that, all method including CFG, have difficulty in depicting some details in the prompts like *fish in its claws* and *eating red berry*. As shown in Figure 15, for failure cases of GFCG where class probabilities of GFCG results are lower than those of CFG, certain features of the species, e.g. white feather of *BALD EAGLE*, are excessively enhanced. This could be caused by improper guidance scales. More visual examples using detailed text prompts are shown in Figure 20 at the end.

5.3. Text-to-image Generation in Pixel Space

We also explored using GFCG in the pixel space alone, without using latent diffusion like SD 1.5 used above. In this test, GFCG is integrated into the DeepFloyd IF model from Stability AI, whose diffusion mechanism is implemented in the pixel level. The IF model is able to generate high definition images in size of 1024×1024 based on given text prompts. We assessed the performance on the Bird Species dataset, and showed some visual examples in Figure 16. The probability of classifying each generated sample as the target class is also included for reference. On multiple bird species, much higher class accuracy was achieved by GFCG over other guidance schemes. The gained accuracy is largely because of enhanced visual qual-

[2] https://github.com/deep-floyd/IF

Figure 10. Visual examples of generated ImageNet class images, comparing GFCG with other guidance methods in EDM2-S sampling. The last column displays examples of the additive method GFCG$_{ATG}$+CFG. While GFCG enhances class accuracy, it sacrifices some diversity. GFCG$_{ATG}$+CFG tries to balance both accuracy and diversity.

ity of the bird, which is more aligned to the actual appearance of corresponding species. Some additional examples are shown in Figure 21 at the end.

6. Limitations and Discussions

6.1. Theoretical Insights into GFCG

The effectiveness of GFCG can be understood through its relation to the score function used in diffusion sampling. In classifier guidance (CG), the denoising process is modified using gradients of the log-probability of the desired class,

$$\nabla_x \log p(c_{\text{des}}|x)\,,$$

which explicitly steers samples toward the target class. However, this requires costly backpropagation.

GFCG instead introduces an *adaptive contrastive mechanism*. By selecting a reference class c_{ref}—typically the most confusing alternative according to the classifier—the guidance direction is approximated as a contrast between the target and reference conditions:

$$D_\theta(x, t, c_{\text{des}}) - D_\theta(x, t, c_{ref})\,.$$

This contrast has two theoretical benefits:

1. **Approximation of classifier gradient.** The difference between classifier scores for the desired and reference classes serves as a finite-difference approximation to the gradient direction used in CG, but without explicit backpropagation.
2. **Adaptive scaling with confidence.** The dynamic guidance weight ω scales updates by how uncertain the classifier is about the current sample. High confidence yields modest corrections, while low confidence amplifies the separation between c_{des} and c_{ref}. This adaptivity prevents over-correction when the sample already aligns with the target, and strengthens guidance when ambiguity is high.

Together, these mechanisms explain why GFCG consistently improves *precision*: it biases the score function toward the desired class in a gradient-free manner while mitigating noise sensitivity through adaptive scaling. Although this introduces a trade-off with diversity, the effect is systematic and complementary to other guidance methods, as our experiments demonstrate.

6.2. Classifier Dependence

While the proposed method is gradient free as it is not subjected to the time consuming gradient calculation, it still inherits other limitations of classifier guidance. First, it requires an existing classifier or a newly trained one for guided sampling, hence the guided sampling is only effective to generate images aligned with classifier training data.

Figure 11. Visual examples of generated ImageNet class images comparing GFCG and GFCG$_{ATG}$ in diversity for EDM2-XXL sampling.

For the Bird Species experiments, as the classifier training dataset consists of close-up shots mostly, the classifier is less effective in guided sampling of bird images of other layouts. We also explored CLIP-based guidance (Table 12) as a domain-agnostic variant but it lags behind in both fidelity and diversity comparing to guided with a domain-specific classifier.

Secondly, similar to CG, the classifier is subjected to accuracy degradation at early sampling stage when the intermediate samples are noisy. The original CG retrains a classifier with noise condition to tackle this challenge at the cost of additional training time. In our work, aspired by more recent training-free classifier guidance works, a multi-step denoising method is included to improve accuracy in c_{ref} and ω, and subsequent sampling quality.

Lastly, performance of the classifier like top-1 accuracy could also be a limiting factor for guidance effectiveness. We have demonstrated though, as shown in Table 9, that the proposed GFCG is not very sensitive to classifier degradation. For example the smaller ResNet-18 has a significantly lower top-1 accuracy comparing to ResNet-101 which is used in the main tests. But when applied for GFCG, it still achieves comparable performance in both FD$_{DINOv2}$ and Precision.

6.3. Limitation of c_{ref}

Beyond errors from noisy images or imperfect classifiers, c_{ref} has other inherent limitations. Our primary goal in this work was to demonstrate the core effectiveness of GFCG with a single adaptive reference class, allowing a fair comparison with existing gradient-free guidance methods of similar complexity. However, in practice an intermediate sample may be confused with multiple classes, making extension to top-k reference classes a natural direction. While a detailed study of the trade-off between computational cost and guidance quality is left for future work, our stochastic selection of c_{ref} already serves as an approximation by incorporating multiple candidates across steps. Its observed improvements over a single deterministic c_{ref} suggest clear potential for multi-reference extensions.

6.4. Failure Cases

Similar to other guided sampling methods, GFCG is also subject to the trade-off between image fidelity and diversity. As shown in the main experiments (Tables 2–3), GFCG is complementary to other methods: when combined with ATG or CFG it recovers diversity while retaining fidelity, yielding better overall trade-offs. However, GFCG may magnify the severity of certain failure case, like the TIT

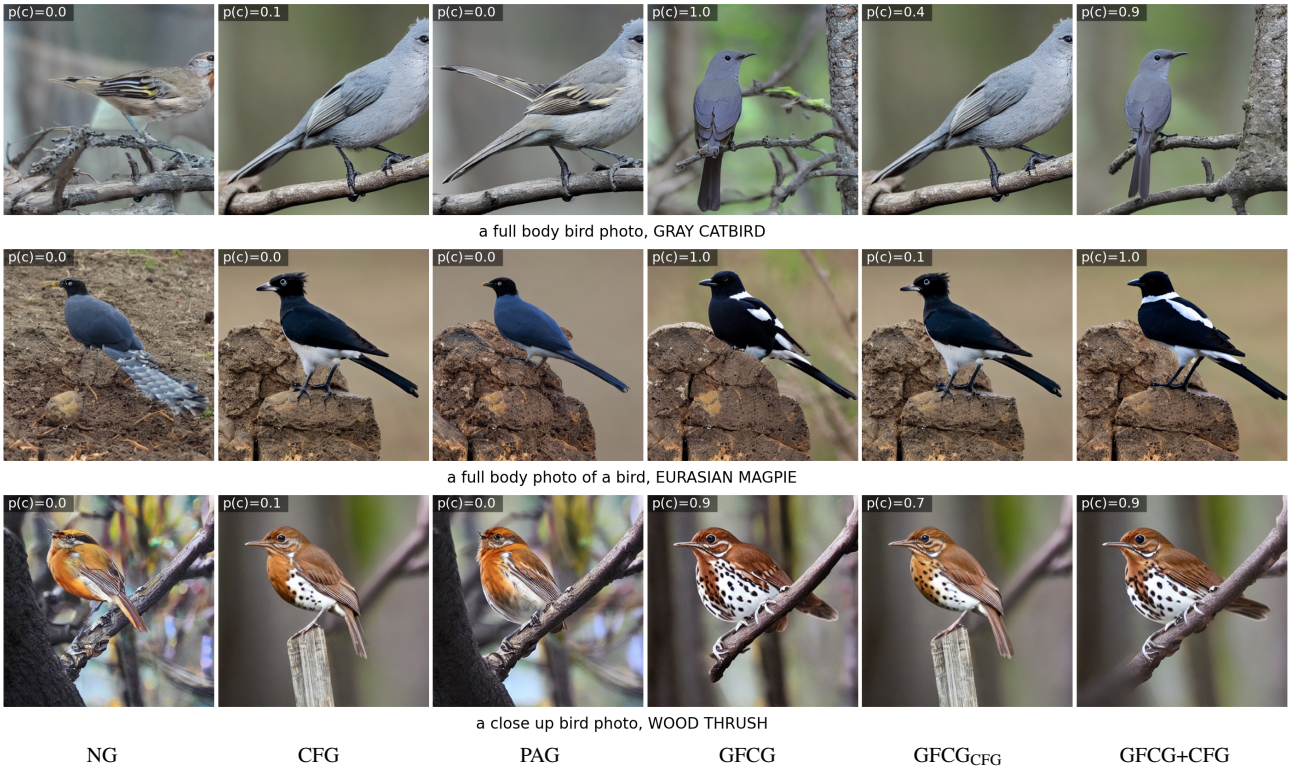


Figure 12. Representative visual examples which demonstrate the benefits of GFCG over others in text-to-image generation using generic text prompts: 1) GFCG and GFCG+CFG improve compositional quality to improve class accuracy; 2) They add the right feather color pattern for higher class probability; 3) GFCG$_{CFG}$ makes minor improvement in chest spot pattern resulting in higher probability.

Figure 13. Representative visual examples where GFCG fails to improve class accuracy using generic prompts, where the incorrect semantic understanding of *MOUSE* and *TEAL* gets enhanced due to GFCG.

MOUSE example shown in Figure 13. The root cause of this failure lies within the text-to-image diffusion model (SD 1.5) as it fails to understand TIT MOUSE as one phrase for a bird species, and future development of the diffusion model can result in elimination of such cases without additional effort. In the meantime, prompt engineering like adding more detailed phrases or negative prompts (like *photo of a mouse*) can help mitigate such risks.

Figure 14. Representative visual examples where GFCG improves class accuracy using detailed text prompts: 1) GFCG enhances the bird features while preserving contextual coherence; 2) GFCG and GFCG+CFG correct the flying posture; 3) or render the right beak color.

References

[1] Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. *arXiv preprint arXiv:2408.00760*, 2024. 3

[2] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35: 26565–26577, 2022. 1, 2

[3] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *arXiv preprint arXiv:2406.02507*, 2024. 1, 3

[4] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24174–24184, 2024. 3

[5] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420, 2022. 1

Figure 15. Representative visual examples where GFCG fails to improve class accuracy using detailed prompts.

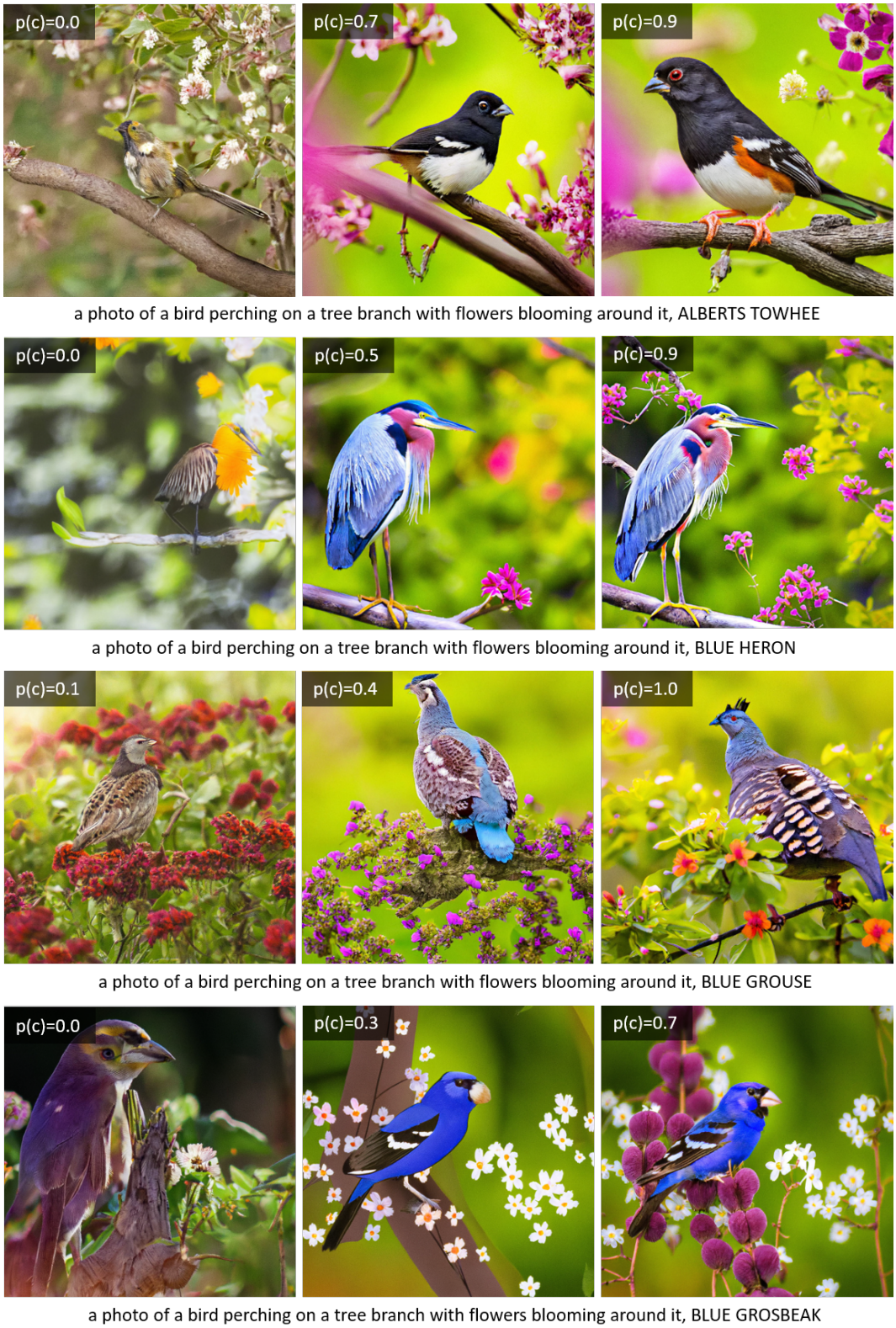


Figure 16. Visual examples which demonstrate the benefits of GFCG in text-to-image generation using pixel-level diffusion model (i.e. without latent diffusion model). The test was performed on the DeepFloyd IF model from Stability AI. GFCG improves the class accuracy of generated image for multiple bird species, by aligning visual presence of the bird to the actual appearance observed in real world.

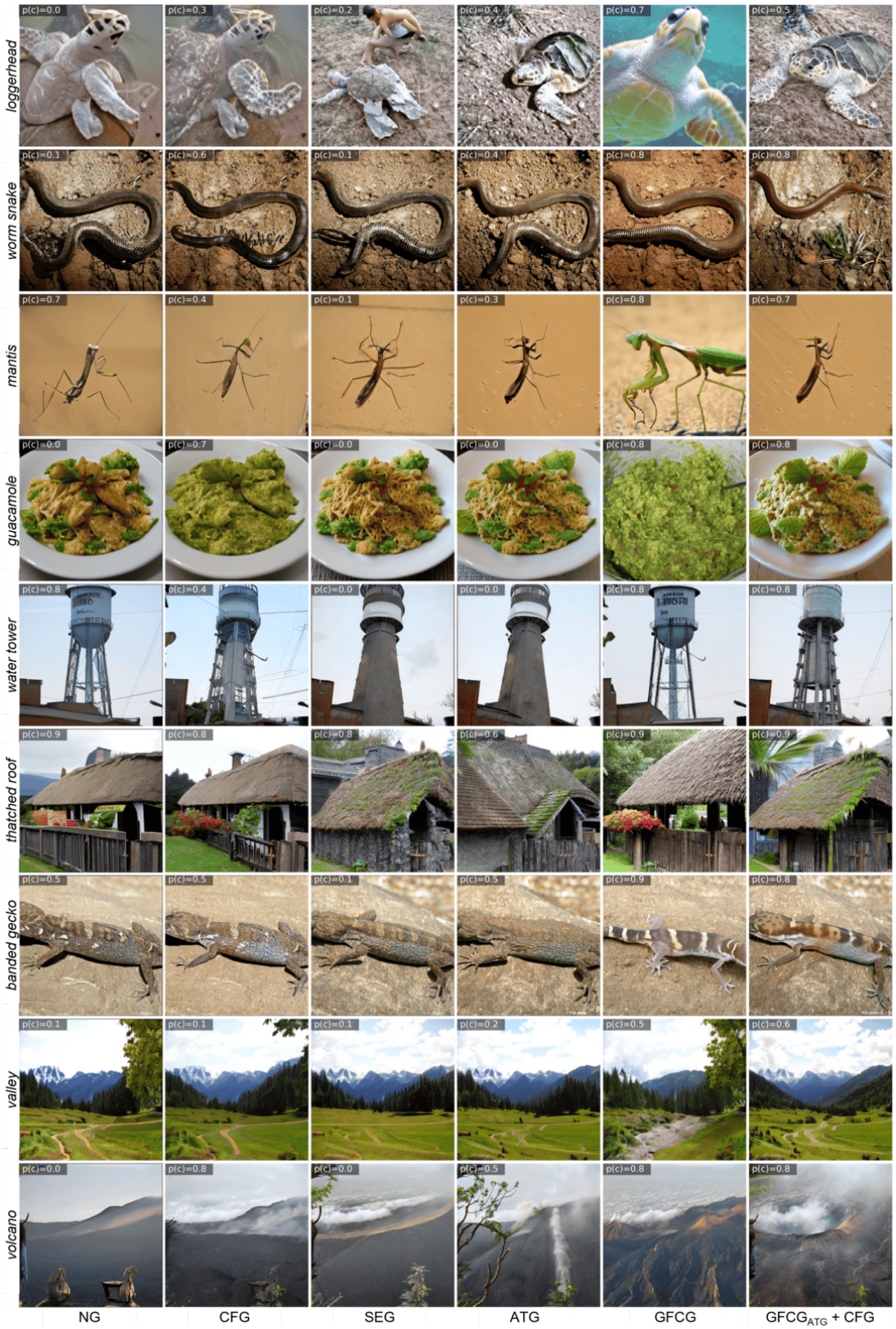


Figure 17. More visual examples of generated ImageNet class images for different guidance methods in EDM2-S sampling.

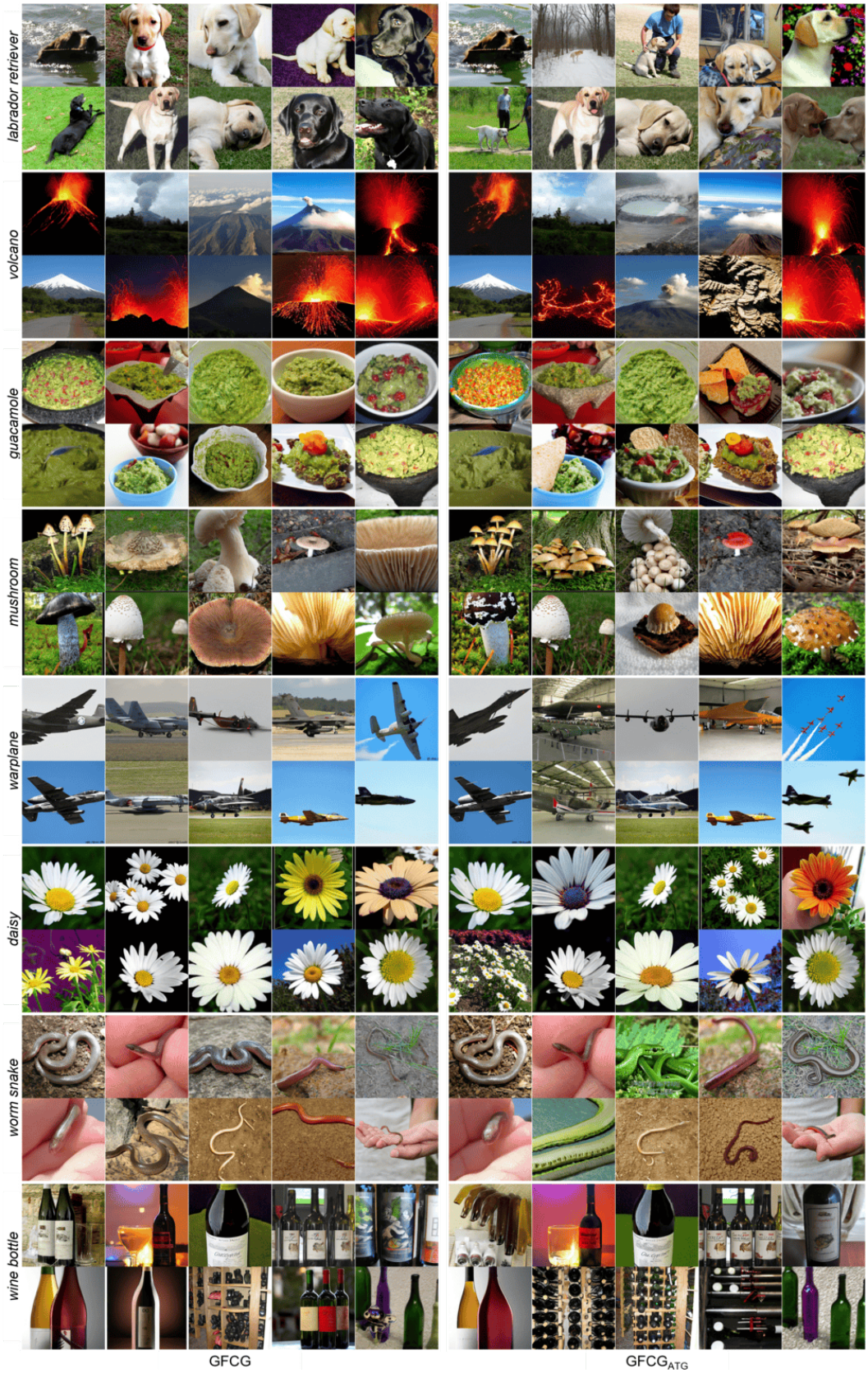


Figure 18. More visual examples of generated ImageNet class images comparing GFCG and GFCG$_{ATG}$ in diversity for EDM2-XXL sampling.

Figure 19. More visual examples from SD 1.5 model using generic text prompts.

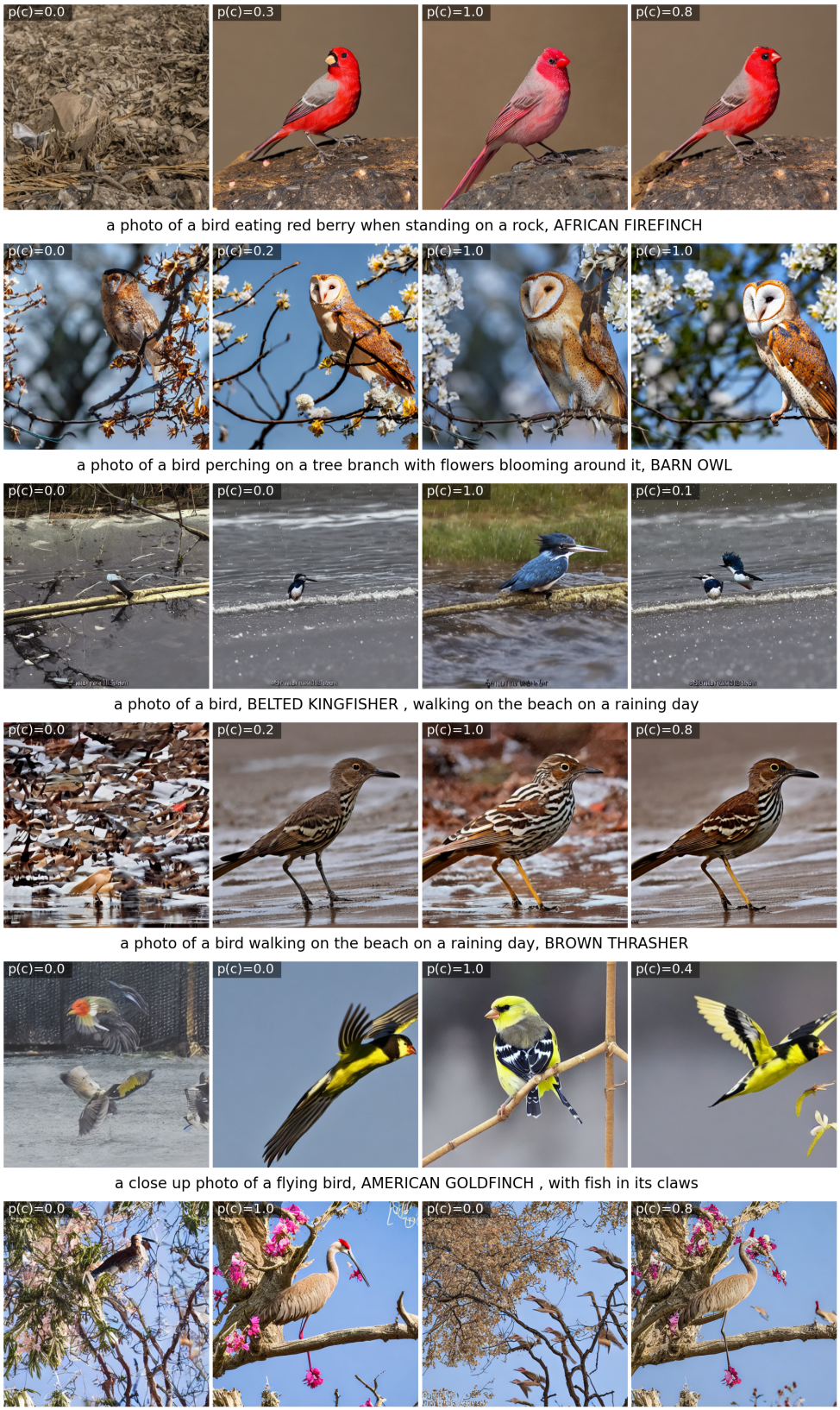


Figure 20. More visual examples from SD 1.5 model using detailed text prompts.

Figure 21. More visual examples from DeepFloyd IF model using detailed text prompts.