A. Outer Steps – Inner Steps Trade-off

The ratio $R_{ND} = \frac{|\text{outer steps}|}{|\text{inner steps}|}$ determines the NFEs required for each update of the Nested Diffusion intermediate prediction. Faster update rates come at the expense of lower quality in the intermediate prediction samples. To illustrate this trade-off, we present Figure 10, which showcases Nested Diffusion sampling with different R_{ND} values while keeping all other hyperparameters and the random seed constant.

To compare the performance of different Nested Diffusion hyperparameter choices, we introduce a novel metric – the Area Under the Curve (AUC) of the log FID per NFE curve. The log FID per NFE curve is defined by the log FID of the images obtained if the algorithm were to be terminated at that particular point in the sampling process. This metric captures the intermediate FID scores, their convergence rate, as well as the frequency of the updates, thus constituting a reasonable metric for anytime generation algorithm evaluation. In the case of Nested Diffusion, the most recent $\hat{\mathbf{x}}_0'$ would be returned until the termination of the first inner diffusion process. From this point, the resulting image would only be updated at the end of each subsequent inner diffusion process. An example of this curve for Nested Diffusion, along with its corresponding AUC, is depicted in Figure 9.

In Table 1, we present a comparison of various R_{ND} ratios for conditional ImageNet [9] generation using our proposed metric. The estimating the log FID per NFE curve is achieved by measuring 50K FID every 10 NFEs for Nested Diffusion totaling 250 NFEs. This metric captures the tradeoff between image quality and update speed, making it relevant for assessing anytime image generation algorithms. We hope this metric proves useful in comparing anytime image generation algorithms in the future.

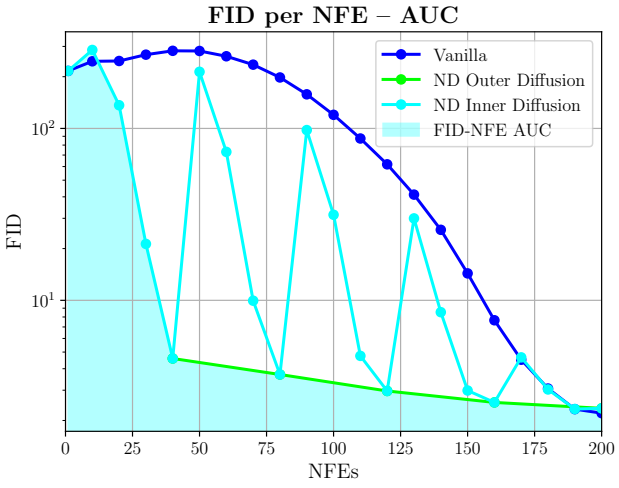

Figure 9. **Graph of the AUC of the FID-NFE curve.**

OUTER STEPS	INNER STEPS	AUC	FINAL FID
1	200	803.33	2.2060
2	100	484.49	**2.1919**
4	50	354.00	2.2677
5	40	**346.15**	2.3534
10	20	388.18	2.6267
20	10	521.30	3.2717

Table 1. **Table of log FID per NFE AUC on ImageNet**. The result reflect different choices of inner steps and outer steps, for a total of 200 NFEs. Vanilla diffusion is equivalent to Nested Diffusion with one outer step, shown in the top line.

B. Anytime Consistency

While maintaining consistency between intermediate samples and the final result is significant for an anytime algorithm, it's equally crucial that the anytime algorithm continues enhancing image quality during the sampling process, leading to incoherence with previous results. Based on these considerations, we conclude that it is desirable to have the semantic details in the generated image remain mostly consistent during anytime sampling, while the image itself may change. Moreover, the user should be made aware of the degree of expected change for each intermediate result produced by the algorithm, should the sampling procedure be continued.

To facilitate a better understanding of the evolution in image dynamics for Nested Diffusion, the average distance of intermediate predictions from the final result is shown in Figure 11, as computed from images generated for the text-to-image experiment in Section 4.2. The following metrics are used; LPIPS [56], image-to-image CLIP Score [15], MSE, and DreamSim [11]. These metrics can give an insight into the consistency dynamics, ranging from non-semantic metrics such as MSE to highly semantic metrics such as DreamSim [11]. From the graphs, we notice that the trend is similar regardless of the choice of R_{ND}. The observed variance for the presented values in Figure 11 is small to negligible.

C. Implementation Details

C.1. Class-Conditional ImageNet Generation

The Dit [37] DNN is trained using Kullback Leibler divergence to yield both the mean and variance of a Gaussian distribution $p_\theta(\mathbf{x}_0|\mathbf{x}_t)$. The model directly predicts the conditional mean of the Gaussian noise in $\mathbf{x}_t$ and the variance of $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$, but we can use a change of variables to view these as the mean and variance of $p_\theta(\mathbf{x}_0|\mathbf{x}_t)$, conforming with our notation. When using the Dit [37] DNN for Nested Diffusion, both the inner diffusion and the outer are conducted in the latent space. The variance prediction

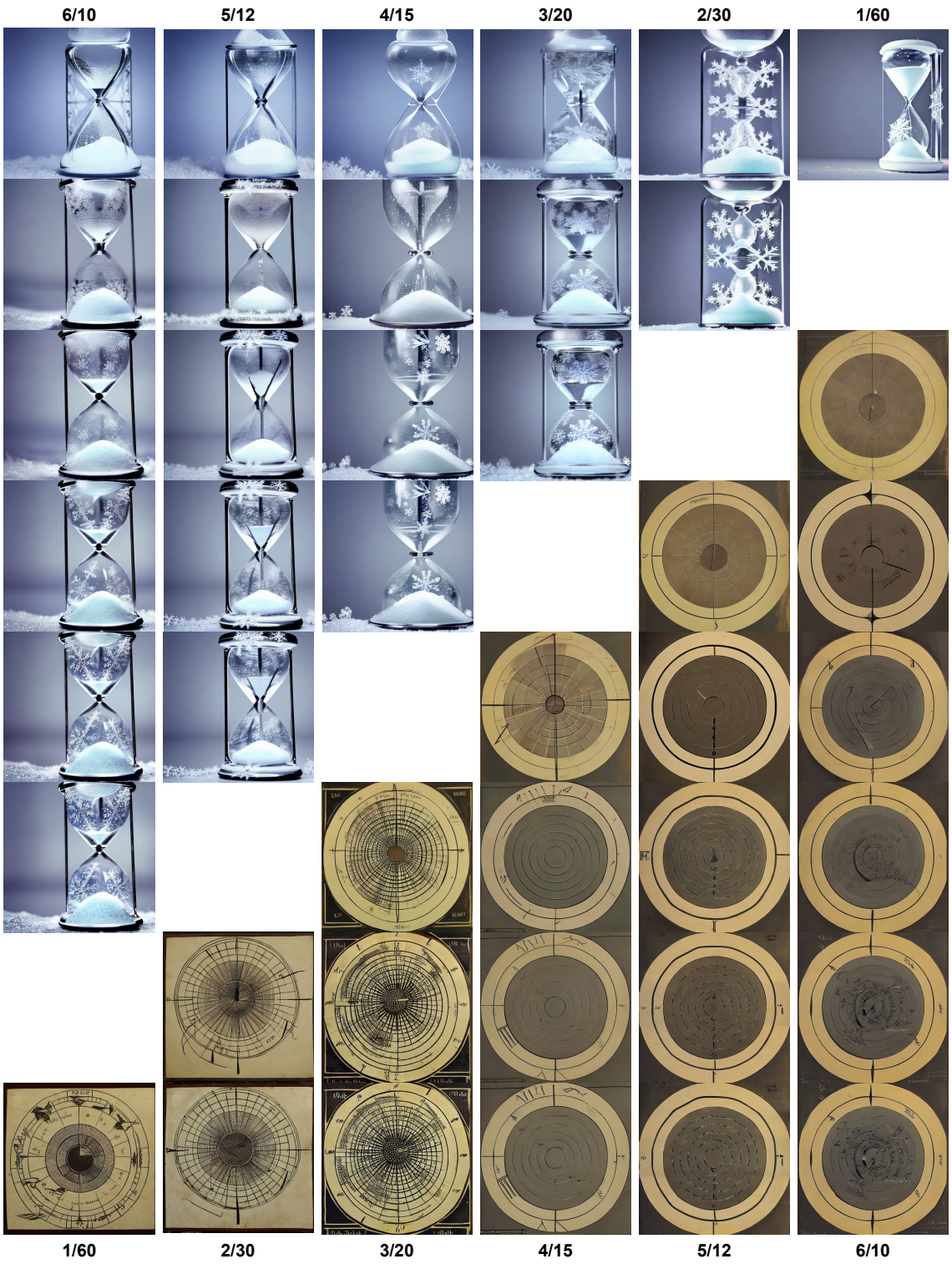

Figure 10. **Qualitative examples of Nested Diffusion with different ratios R_{ND}.** Each column denoted with |outer steps|/|inner steps| at the top or bottom. Top text: *a photograph of an hourglass filled with snowflakes*. Bottom text: *a diagram of an ancient sundial*. Diffusion process progresses from top to bottom.

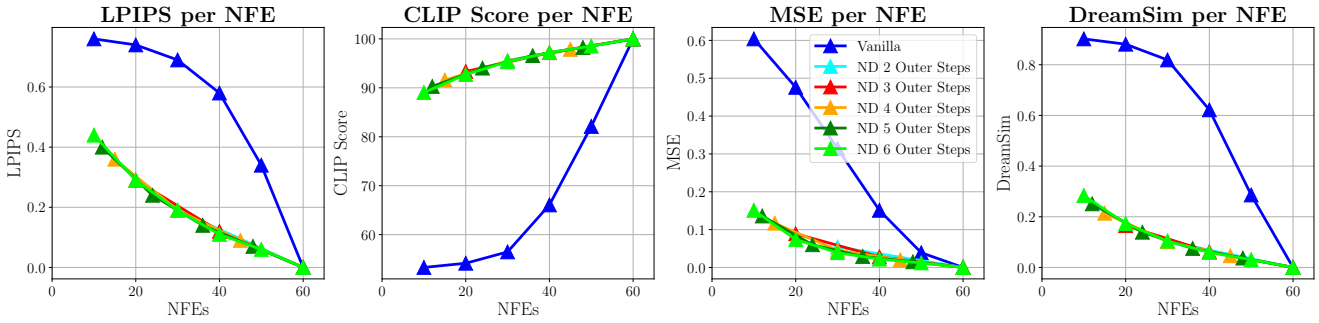

Figure 11. **Progression of distance of intermediate predictions from the final result.** Metrics are (left to right): LPIPS, image-to-image CLIP Score, MSE, and DreamSim.

is used only in the inner diffusion, while the outer diffusion remains deterministic DDIM [48] sampling. CFG [18] is regarded as part of the DNN, and therefore applied in the inner diffusion only. We set the CFG value to 1.5 similar to Peebles & Xie, 2022 [37].

C.2. Text-to-Image Generation

We use Stable Diffusion V1.5 [41] FP32 to generate 512×512-pixel images. We implement Nested Diffusion using non-deterministic DDIM [48] with $\eta = 0.85$ for the inner diffusion, and treat the CFG [18] as we did in Section 4.1, setting it to the default value of 7.5. No clipping or thresholding is applied, and final $\bar{\alpha}_t$ set to zero. Due to the size limit for submission, the images shown in the paper and supplementary material have been compressed using JPEG, which may impact the perceptual image quality.

In the MS-COCO [29] FID [16] evaluation we follow the protocol in [2,40,41,43], using a budget of 60 NFEs per image and using the FP16 version of Stable Diffusion V1.5 for all configurations. All other hyperparameters remain as specified above.

The high-order solver (DPM-Solver++ [33]) setup used the hyper parameters from above except for the following; The inner diffusion was based on DPM-Solver++(2S) [33], with default hyperparamers. The outer diffusion was changed to DDIM with $\eta = \sqrt{1-\bar{\alpha}_t}$, for larger stochasticity.

In addition to the MS-COCO FID shown in Figure 3b, we present the average CLIP Score [15] of the generated images with their guidance prompt in Figure 12. The CLIP Score results show a similar trend to their FID counter parts – Nested Diffusion achieves a high score on the intermediate results and a slightly improved final image result compared to vanilla diffusion.

C.3. Inverse Problem solving on CelebA-HQ256

We evaluate the following inverse problem tasks; denoising of additive white Gaussian noise with variance set to 1.0, block-super-resolution with factor 16, colorization, and inpainting of 50% random pixels in the image. More information on these degredations can be found in DDRM [22].

Algorithm 3 Inverse Problem Solving using Nested Diffusion

Outer diffusion denoted in blue, with step size s^o
Inner diffusion denoted in purple, with step sizes $\{s_t^i\}$
$\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$
for t in $\{T, T-s^o, \ldots, 1+s^o, 1\}$ **do**
  $\mathbf{x}'_t = \mathbf{x}_t$ ▷ Beginning of inner diffusion
  for τ in $\{t, t-s_t^i, \ldots, 1+s_t^i, 1\}$ **do**
    $\hat{\mathbf{x}}'_0 \sim p_\theta(\mathbf{x}'_0 | \mathbf{x}'_\tau, \mathbf{y})$
    $\mathbf{x}'_{\tau - s_t^i} \sim q'(\mathbf{x}'_{\tau-s_t^i} | \hat{\mathbf{x}}'_0, \mathbf{x}'_\tau)$
  end for
  $\hat{\mathbf{x}}_0 = \mathbf{x}'_0$ ▷ End of inner diffusion
  $\mathbf{x}_{t-s^o} \sim q(\mathbf{x}_{t-s^o} | \hat{\mathbf{x}}_0, \mathbf{x}_t)$
end for
return $\mathbf{x}_0$

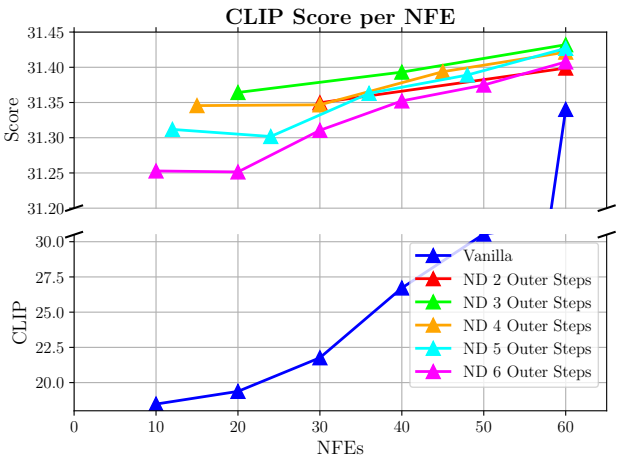

Figure 12. **Average CLIP Scores of images generated with DDIM inner diffusion process.**

PSNR↑	10 NFEs	20 NFEs	30 NFEs	DDRM
DENOISING	25.76	25.80	25.81	25.83
SR16	23.19	23.55	23.89	23.78
COLOR	19.47	21.54	21.55	23.92
INPAINTING	31.20	33.16	35.07	35.08
FID↓	**10 NFEs**	**20 NFEs**	**30 NFEs**	**DDRM**
DENOISING	19.11	16.70	12.97	12.24
SR16	16.77	13.79	11.65	11.30
COLOR	12.78	7.08	6.95	4.28
INPAINTING	14.63	7.57	3.26	3.18

Table 2. **PSNR and 30K FID of inverse problems solving on CelebA-HQ256.** The inverse problems include denoising of additive white Gaussian noise, block super-resolution with a factor of 16, colorization, and inpainting of random pixels, listed from top to bottom.

our Nested Diffusion examples all use default η hyperparamters.

The inverse problem solving algorithm using Nested Diffusion is shown in Algorithm 3. The inner diffusion is composed of a complete inverse problem sampling process (notice the similarity to Algorithm 2). In our experiment, we have used DDRM [22], an iterative sampling process, as the aforementioned inverse problem sampling process.

Table 2 presents PSNR and FID evaluations for Nested Diffusion on inverse problem solving. The metrics were generated on 30K samples from the CelebA-HQ256 [21] dataset. We note that Nested Diffusion's final results are comparable to vanilla DDRM.

D. More Examples

We provide more examples for images generated from various experiments below.

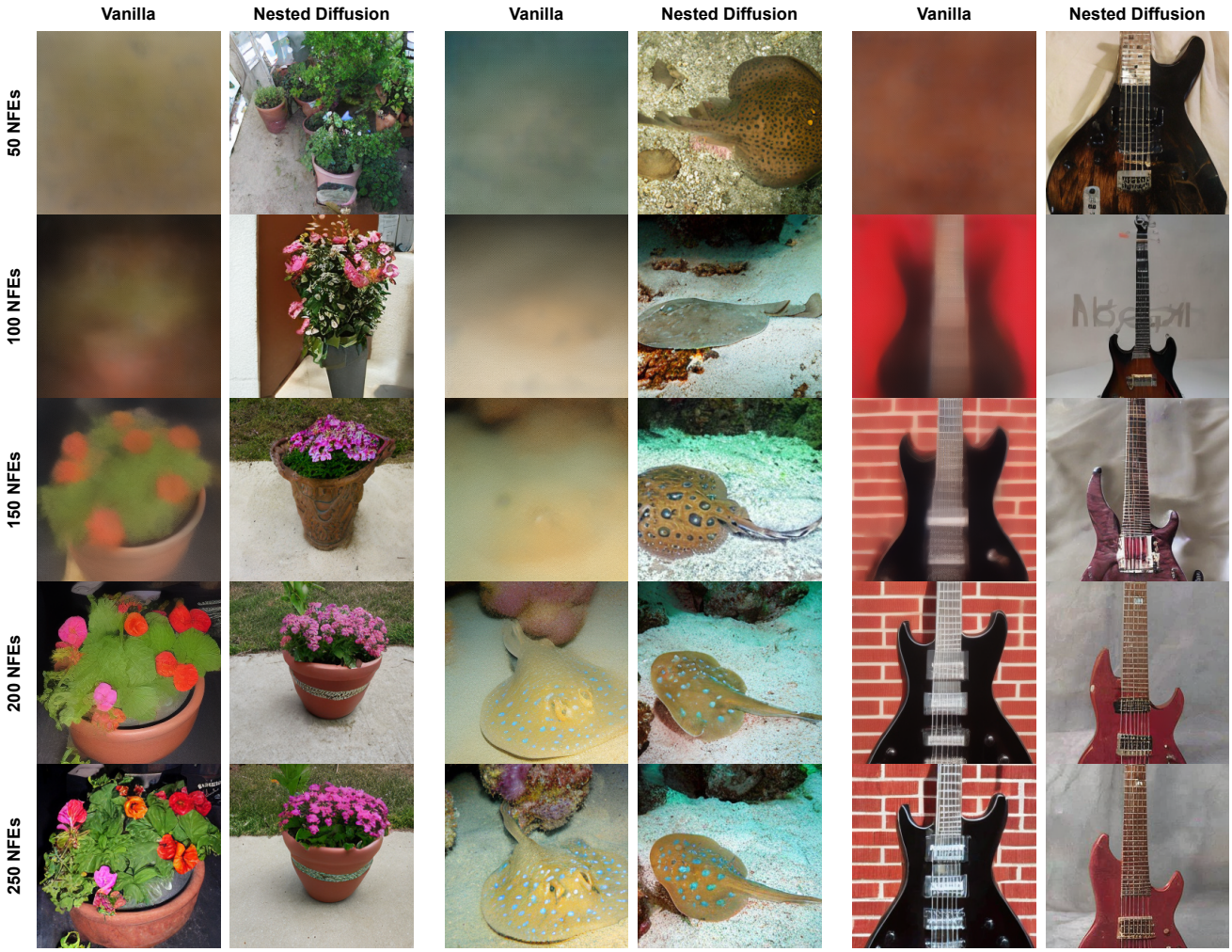

Figure 13. **Additional samples of ImageNet generation, comparing vanilla diffusion model to Nested Diffusion.**

	TOTAL 100 NFES			TOTAL 150 NFES			TOTAL 200 NFES			TOTAL 250 NFES		
%	NFES	VAN	ND	NFES	VAN	ND	NFES	VAN	ND	NFES	VAN	ND
20%	20	282.89	13.03	30	282.05	6.57	40	282.83	4.58	50	284.13	3.57
40%	40	202.34	9.20	60	199.74	4.99	80	197.93	3.70	100	197.74	3.08
60%	60	65.22	5.97	90	62.37	3.58	120	61.82	2.96	150	60.19	2.61
80%	80	8.10	4.00	120	7.67	2.82	160	7.65	2.54	200	7.57	2.36
100%	100	2.44	3.18	150	2.24	2.50	200	2.20	2.35	250	2.16	2.28

Table 3. **Exact 50K FID evaluation of Nested (ND) and vanilla (Van) diffusion processes**. The intermediate prediction are measure when stopped at different percentages of the full algorithm runtime (100, 150, 200, 250 NFEs).

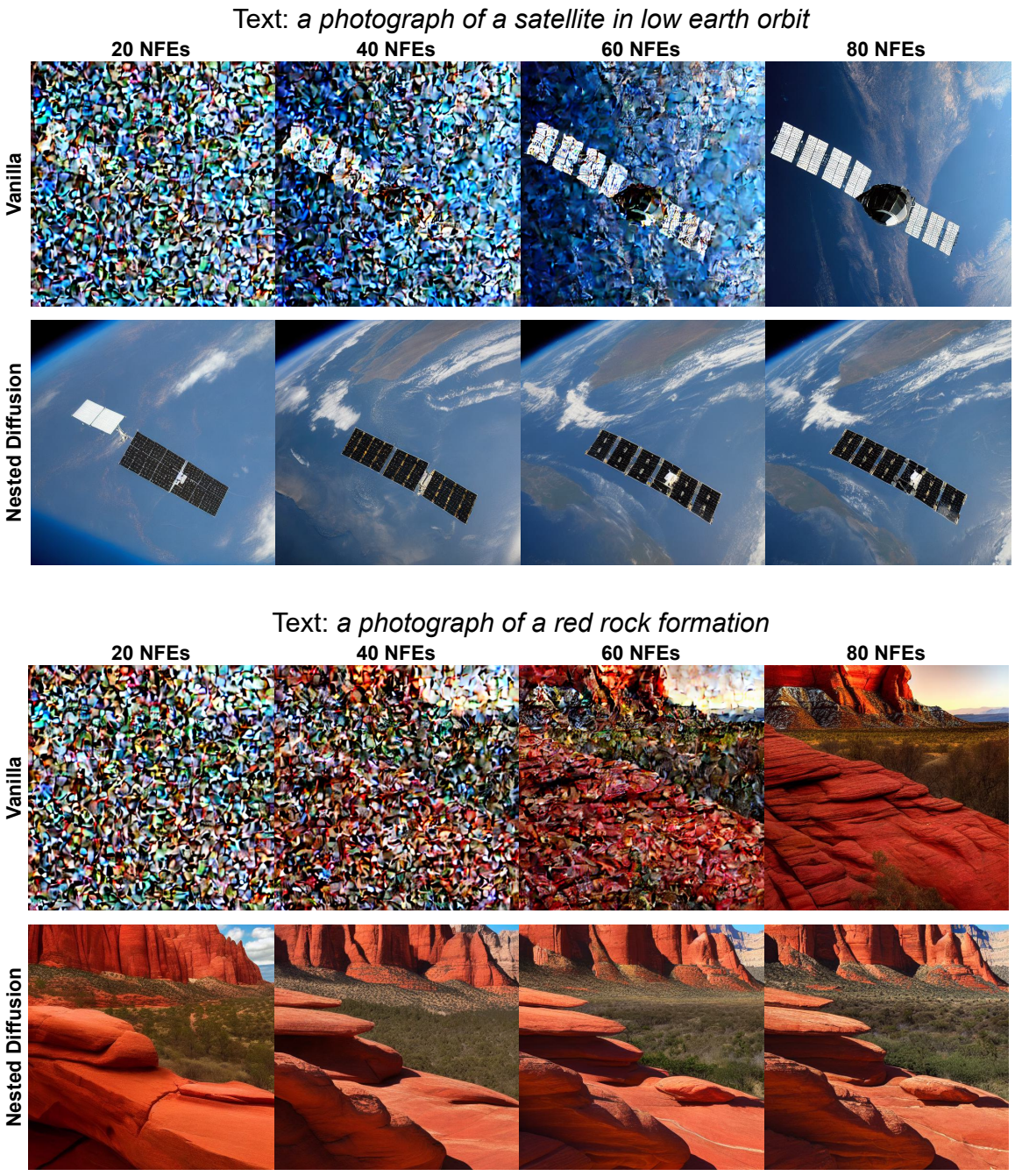

Figure 14. **More Results of intermediate predictions of Stable Diffusion from a reverse diffusion process with 80 steps.**

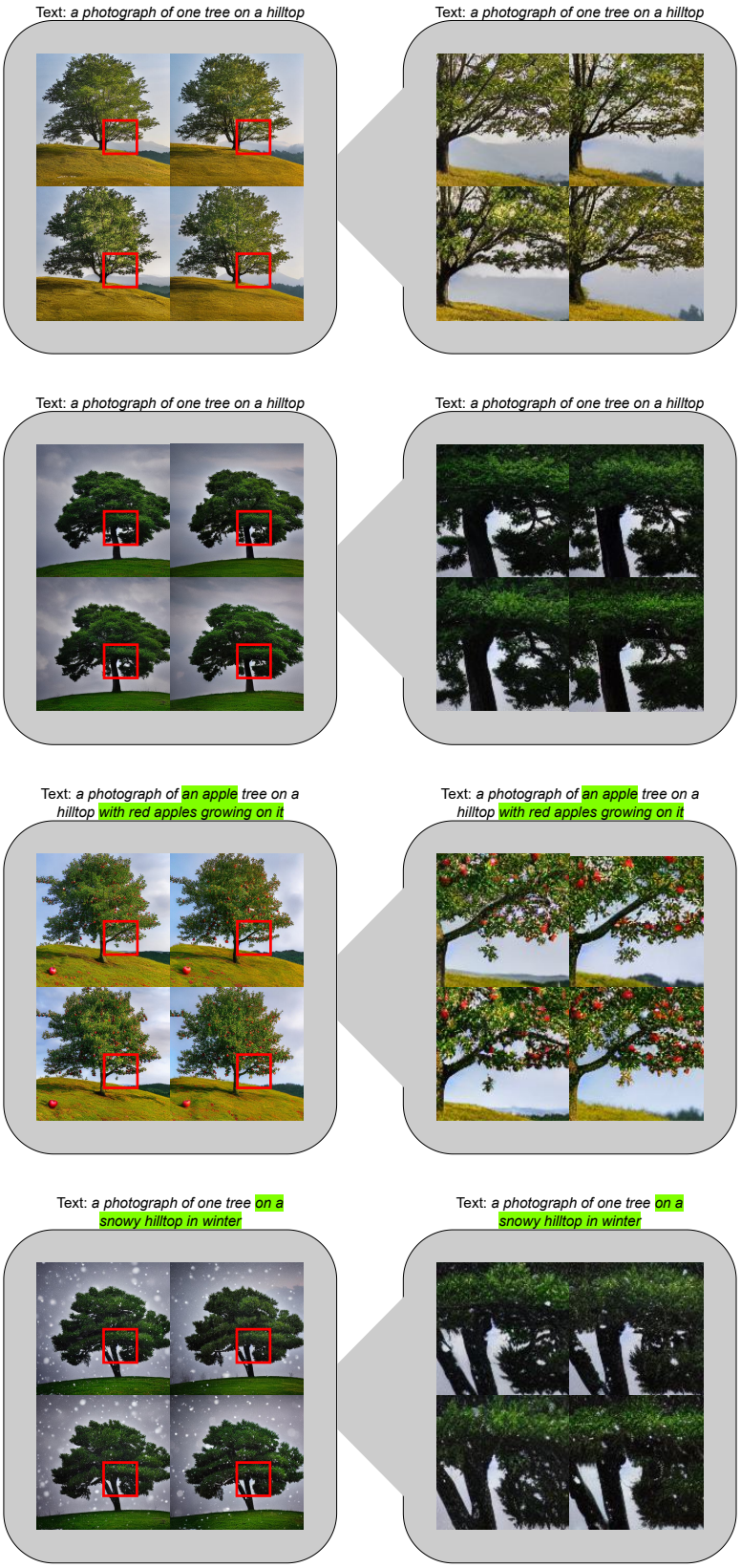

Figure 15. **Zoom-in on the final images of Figure 7 and Figure 8 for viewing fine details in the images.**