Guidance Scale Effects in Classifier-Free Diffusion Models

Matt McWilliams

December 2025

1 Introduction

In this experiment, we study a conditional diffusion model trained on the MNIST dataset, analyze the efficacy of classifier-free guidance, and explore the effects of the guidance scale on the clarity and diversity of generated samples. We then train a simple classifier on MNIST and analyze guidance strength's effects on classifier accuracy of generated samples. This writeup does not propose a new method; rather, it clarifies the distributional effects and failure modes of the guidance scale.

2 Setup

The MNIST digits are grayscale 28×28 images with labels $y \in \{0, \ldots, 9\}$, normalized to $[-1, 1]$.

The forward process used exclusively during training is a Markov process, which begins with the data distribution x_0 and iteratively adds time-dependent Gaussian noise to each digit over T time steps, causing the distribution to lose structure and tend towards random noise. The forward process q is defined as

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; x_{t-1}\sqrt{1-\beta_t}, \mathbf{I}\beta_t), \tag{1}$$

where β_t is the amount of noise added at time t, sampled from a cosine beta schedule. The forward schedule is re-parameterized to enable accelerated training:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}. \tag{2}$$

$$\bar{\alpha}_t = \prod_{k=0}^{t} \alpha_k \qquad \alpha_t = 1 - \beta_t \tag{3}$$

Written differently, the forward process can be sampled at any timestep t by linearly combining the initial distribution x_0 with noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ using the following formula:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + (1-\bar{\alpha}_t)\epsilon \tag{4}$$

The reverse process is learned in training and used in sampling. A time-conditioned model m with parameters θ is trained to predict the ϵ used in the forward process using MSE loss:

$$\mathcal{L}_\theta = \mathbb{E}_{x_0,\epsilon,t}[\|\epsilon - m_\theta(x_t, t, y)\|^2]. \tag{5}$$

The model is conditioned over t and y. During sampling, the model iteratively removes noise using the following formula and returns x_0 as a generated sample. We use m_θ to denote network output and ϵ_θ to denote the effective noise estimate used during sampling.

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t, y)) + \sigma_t z \tag{6}$$

Because we condition the model on labels, rather than training a new model for every label, setting $\epsilon_\theta(x, t, y) := m_\theta(x, t, y)$ results in insufficient separation between class-conditioned modes. To prevent this, during training, label dropout is used to train the model both conditionally and unconditionally. ϵ_θ is then defined as a linear combination of the conditional and unconditional predictions:

$$\epsilon_\theta(x, t, y) := (1 + s)m_\theta(x, t, y) - sm_\theta(x, t, \varnothing), \tag{7}$$

where $s \geq 0$ is the guidance scale. Doing this allows the model to amplify label-dependent features and biases the reverse process towards higher conditional likelihood.

3 Method

In these experiments, a UNet with 16 base channels, residual-block layers, two max-pool downscale layers, and two learned upscale channels were used. The model is conditioned over the timestep t and the label y. t was encoded using sinusoidal embedding and added to the hidden layers within each residual block. y was similarly added to each hidden layer using a learned embedding. The model was conditioned over $T = 1000$ timesteps.

The model was trained over 4000 gradient steps using Adam with learning rate 10^{-3}, batch size 128, and label dropout probability $p = 0.1$.

The model is sampled using stochastic DDPM noise prediction described in equation 6. ϵ_θ is calculated using classifier-free guidance: a linear combination of the conditional model prediction and the unconditional model prediction dependent on the guidance strength s, described in equation 7.

Classifier-free guidance operates under the empirical observation that the conditioned model does not adequately emphasize the label-dependent features and assumes that the difference between conditional and unconditional estimates estimates a class-conditioned score component.

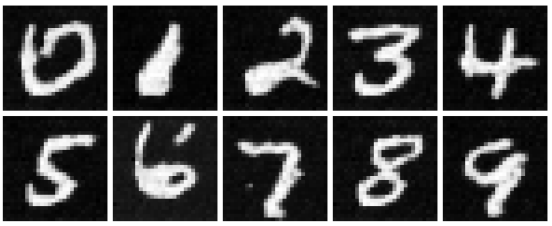

Figure 1: Sample of each digit; guidance strength $s = 3$.

4 Experiments

4.1 Class-Controlled Samples

First, we sample the conditioned model with classifier-free guidance to test if it could reliably generate each digit. A guidance strength $s = 3$ is used. As shown in Figure 1, each digit class is visually distinguishable and the samples shown do not contain obvious cross-class artifacts.

4.2 Guidance Scale Samples

Next, we sample the class $y = 3$ at each guidance strength to test the effects of the guidance strength s on the diversity and correctness of samples. The random seed used during sampling is fixed such that the seed is the same for each guidance strength. This makes discrepancies between each trial dependent only on the guidance strength, but it can bias the distributional measurements if the random seed generates an improbable distribution.

As shown in Figure 2, low guidance strengths ($s \leq 1.0$) result in samples not representative of the distribution of class $y = 3$, with artifacts from the class $y = 8$ particularly visible in the shown samples. Guidance strength $s = 0.0$ represents the conditioned model without any classifier-free guidance and results in samples that are not distinguishably a 3. High guidance strengths ($s \geq 15.0$), on the other hand, result in samples with extreme values, exaggerated scale of label-dependent features, and lack of diversity. For the sampling setup of the digit 3 used above, a favorable tradeoff with minimal artifacts but clear diversity is the range $s \in [2.0, 5.0]$.

4.3 Classifier Evaluation

Finally, to produce quantitative evidence about guidance scale parameterization, we trained a classifier to predict a digit's label from the sampled pixels and analyzed the classifier's accuracy and confidence by guidance scale.

The classifier c with parameters ϕ has two convolutional layers that map to 4 channels separated by a max-pooling layer of kernel size 2, followed by two linear layers with 16 hidden neurons, ReLU activation function, and 10 output neurons that pass through softmax to produce probabilities. The classifier was

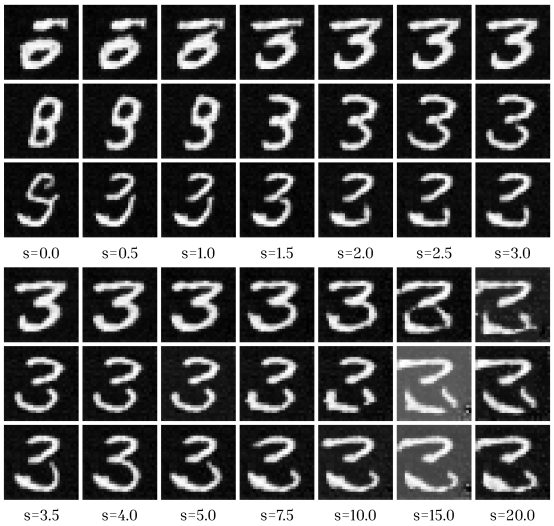

Figure 2: Three samples of the digit 3, generated at 14 different guidance scales. From left to right: $s = 0.0$, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 5.0, 7.5, 10.0, 15.0, 20.0. Note that the lighter backgrounds of $s = 15$ originate from rescaling the image values; the samples had extreme negatives, not lighter backgrounds.

trained over real MNIST only, not generated samples, with 4000 iterations using Adam optimizer with batch size 128 and learning rate 10^{-3}. During training, a small amount of noise ($w \sim \mathcal{N}(0, 0.1)$) was added to each input to account for the generated samples' small amount of background noise. The loss function $\mathcal{L}_\phi$ is the MSE between the one-hot embedding of y and the predicted class probability vector, $c_\phi = \text{softmax}(f_\phi)$, with $C = 10$ representing the number of classes. Although cross-entropy is standard for classification, we use mean squared error for simplicity and because relative accuracy across guidance scales, not absolute calibration, is useful for our comparative analysis.

$$\mathcal{L}_\phi = \frac{1}{C} \sum_{i=1}^{C} (e_{y,i} - c_{\phi,i})^2 \tag{8}$$

The diffusion model was used to generate 200 samples of random digits at each guidance scale. For reasons described in Section 4.2, the random seed was fixed across guidance scales. After each sample is generated, the classifier is run to assign a probability of the correct label y being assigned to the digit, as shown in Figure 3. Statistics of these probabilities are collected in Table 1. The model generates samples with classifier accuracy greater than 90% when $0.75 \leq s \leq 3.5$, with the highest accuracy at $s = 2.0$. Assuming classifier confidence is highest when the samples are most similar to the MNIST training data, this suggests that, under the provided sampling conditions, the diffusion model generates samples most similar to the training data when $s = 2.0$, although

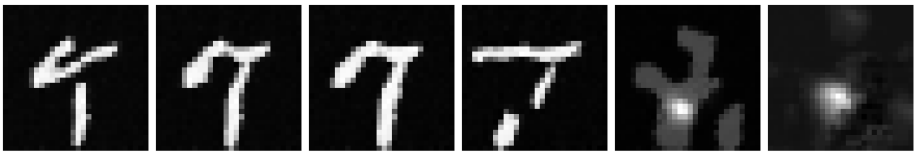

Figure 3: Noise-controlled samples of the digit $y = 7$ at various guidance strengths $s \in \{0, 1, 3, 5, 7.5, 10\}$, and the probability p assigned by the classifier of the digit $y = 7$ for each sample. Contrary to the digit $y = 3$, shown in Figure 2, which consistently generated high-quality samples even at higher guidance scales ($s \geq 7.5$), the sample quality of the digit $y = 7$ begins to deteriorate as low as $s = 5$.

s	μ_p	σ_p	Acc_y	$[0, 0.01)$	$[0.01, 0.1)$	$[0.1, 0.9)$	$[0.9, 1.0]$
0.00	0.6948	0.4055	0.700	0.090	0.120	0.195	0.595
0.25	0.7792	0.3566	0.785	0.055	0.070	0.210	0.665
0.50	0.8501	0.3040	0.850	0.0350	0.040	0.135	0.790
0.75	0.8983	0.2569	0.915	0.0300	0.025	0.105	0.840
1.00	0.9215	0.2197	0.935	0.025	0.010	0.105	0.860
1.25	0.9465	0.1825	0.965	0.020	0.010	0.065	0.905
1.50	0.9531	0.1763	0.965	0.020	0.005	0.060	0.915
1.75	0.9591	0.1594	0.965	0.015	0.000	0.065	0.920
2.00	0.9565	0.1668	0.970	0.015	0.005	0.055	0.925
2.50	0.9469	0.1960	0.955	0.025	0.005	0.045	0.925
3.00	0.9340	0.2239	0.940	0.040	0.005	0.045	0.910
3.50	0.9026	0.2711	0.905	0.065	0.010	0.060	0.865
4.00	0.8561	0.3318	0.855	0.090	0.025	0.055	0.830
5.00	0.8325	0.3609	0.840	0.125	0.025	0.035	0.815
7.50	0.6582	0.4611	0.655	0.290	0.015	0.065	0.630
10.0	0.3821	0.4681	0.385	0.545	0.040	0.075	0.340
15.0	0.0985	0.2827	0.105	0.840	0.045	0.035	0.080
20.0	0.1512	0.3462	0.150	0.795	0.035	0.040	0.130

Table 1: Classifier accuracy on various guidance scales, with sample size at each guidance scale $n = 200$, guidance scale s, mean probability assigned to correct label by classifier μ_p, standard deviation of probability σ_p, and top-1 accuracy Acc_y. Each range indicates the fraction of samples with assigned probabilities < 0.01, $0.01 - 0.1$, $0.1 - 0.9$, and > 0.9.

samples are still easily identifiable for $0.75 \leq s \leq 3.5$.

5 Conclusion

We examined how classifier-free guidance affects the sample quality and diversity of conditioned diffusion models on the MNIST dataset. We found that lower guidance scales tend to result in more diversity but also more cross-class artifacts, while higher guidance scales tend to result in less diversity, extreme values, and exaggerated label-dependent features. Additionally, having fixed noise and fixed digit sampling may bias our training; because different digits have different optimal guidance scale, randomly sampling digits could result in imbalanced experimentation, even at sample sizes as high as $n = 200$. Finally, the MNIST dataset is much simpler than other datasets and thus may exhibit different behavior than datasets with more complicated distributions. A key observation we made was that for some digits, such as $y = 3$, samples were still high quality even at higher guidance scales ($s = 10.0$), while other digits' sample qualities, for digits such as $y = 7$, deteriorate at much lower guidance scales ($s > 5.0$), suggesting that a single global guidance scale may be suboptimal even for simple datasets.