GeoGuide: Geometric guidance of diffusion models

Mateusz Poleski [1] Jacek Tabor[1] Przemysław Spurek[1,2]
[1]Faculty of Mathematics and Computer Science, Jagiellonian University
[2]IDEAS NCBR
mateusz.poleski@student.uj.edu.pl {jacek.tabor, przemyslaw.spurek}@uj.edu.pl

Abstract

Diffusion models have emerged as powerful tools for image generation, offering flexibility in generating images conditioned on specific classes or properties. Unlike GANs, diffusion models can be conditioned during training with relative ease.

However, adapting pre-trained diffusion models to generate images from new, unlabeled data remains a significant challenge. The ADM-G approach addresses this by guiding diffusion models to generate images from a given class, but it often produces results of lower quality compared to models originally trained with class-specific conditioning. For instance, the ADM-G-guided model achieves an FID score nearly three times worse than that of a class-conditioned guidance. We identify that this performance gap arises partly because ADM-G provides minimal guidance during the final stages of the denoising process. To overcome this limitation, we introduce GeoGuide, a novel guidance method that improves the model's trajectory alignment with the data manifold. GeoGuide refines the backward denoising process by applying normalized adjustments to the model's output. Experimental results show that GeoGuide significantly outperforms ADM-G in both FID scores and the visual quality of the generated images.

1. Introduction

Diffusion models are crucial for generating images with specified characteristics. Compared to GAN models, their benefit is that they can be easily conditioned to produce outcomes with the desired attributes [1, 3, 9].

However, given a pre-trained diffusion model, it is not trivial to construct images that belong to the class that was not considered earlier in the conditioning process. One of the possible solutions is given by the guidance model ADM-G [3]. Roughly speaking, to produce an image from a new class, we first train a classifier for this class and then add the rescaled gradient during the backward process. ADM-G introduces guidance through probabilistic principles, for a more detailed description, see Section 4. Unfortunately, ADM-G shows a notable difference in the FID score between the model that only uses guided techniques and the model with class-conditioned guidance. Specifically, the ADM-G model achieves a 4.59 FID score with class-conditioned guidance, while the guided model obtains only 12 FID score, as referenced in Table 4 of [3].

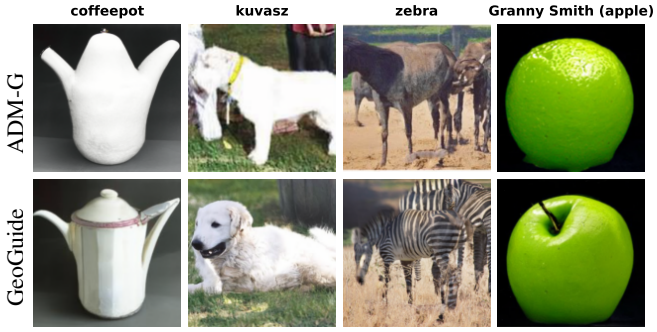


Figure 1. In our paper, we propose a new approach to guidance of diffusion models, called GeoGuide. In contrast to ADM-G [3] we use updates with the same norm and consequently keep the guided diffusion process close to the manifold of a given class. Observe that this allows us to construct images with more details characteristic of a given class, resulting in a decrease in the FID score from 12 in ADM-G to 7.32 in GeoGuide, see Table 2. The images were constructed with the same diffusion noise for ADM-G and GeoGuide.

This paper aims to reduce the gap between the quality of images generated by diffusion models guided on previously unlabeled data compared to class-conditioned guides, see Figure 1. To do so, we switch perspectives from probabilistic to metric. We postulate that the trajectories of the guided model should be close to the metric properties of the unguided ones. It occurs that by applying ADM-G guidance, the norms of adjustment in to the backward diffusion denoising decrease significantly in the last iterations of the denoising procedure, see Figure 2. As a result, the images become nearly unguided at half of their trajectory and con-

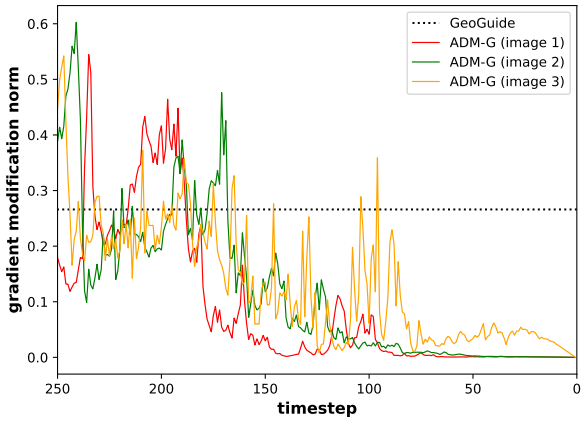


Figure 2. Norm values of the gradient modification factor applied at each iteration of the classifier guided diffusion sampling backward process. Comparison of image generated with GeoGuide and three random images generated with ADM-G. Observe that in the case of the vanilla guidance (probabilistic approach) the norm of the modification at the last steps of the diffusion process is close to zero, which results in less detail in the produced images, see Figure 1.

sequently lose the ability to produce details specific to the given class, see Figure 1 and Figure 3.

To address this issue, we shift the perspective from a probabilistic framework to a metric-based method. We theoretically demonstrate that the trajectory of the diffusion model lies close to the data manifold. Using such intuition, we propose a new guidance model GeoGuide, which uses fixed length updates to force the denoising process to be as close as possible to the data manifold. GeoGuide use norm of classification gradient to control updates. Therefore, our model is easy to implement and outperforms the probabilistic approach with respect to FID score and the quality of generated images.

Concluding, the main contributions of the paper are the following:

- we propose a new guidance model GeoGuide, motivated by the metric properties of the trajectories of the diffusion process,
- GeoGuide is easy to implement and controls the norm of the guidance,
- GeoGuide outperforms ADM-G in pure guidance with respect to FID score and quality of generated images.

2. Related Works

Our method aims to improve sample generation of pretrained diffusion models [3, 13, 14]. It can be easily incorporated into existing models as it does not modify the structure or inner workings of the network [2, 7]. It only modifies sampling logic, which allows one to easily leverage the power of already existing diffusion models and classifiers. In [5] authors present another approach to enhance sampling quality in diffusion models. They defined alternative way of training for time-dependent adversarially robust classifier, and use it as guidance for a generative diffusion model. These classifier gradients are better aligned with human perception, and could better guide a generative process towards semantically meaningful images. In Section 5 we tried to combine robust classifier with our GeoGuide and evaluated results.

Another interesting technique for improving guidance in score-based diffusion models is Discriminator Guidance [6]. The authors propose integrating a discriminator, commonly used in Generative Adversarial Networks (GANs), to guide the generative process. The key idea is to use the discriminator to evaluate and refine the intermediate states of the generative process, improving overall sample quality. Integrating discriminator guidance helps mitigate mode collapse and improves sample diversity and fidelity. This novel approach combines discriminator and vanilla classifier guidance in the generation process. It is possible that integrating GeoGuide into this approach, instead of a vanilla guidance, could lead to even better results.

As we can see in Figure 5, diffusion models, especially when used with high guidance scale values to achieve optimal image quality, are prone to limited output diversity. One potential solution to this problem is the Condition-Annealed Diffusion Sampler (CADS) [10]. In this approach, the sampling strategy anneals the conditioning signal by adding scheduled, monotonically decreasing Gaussian noise to the conditioning vector during inference to balance diversity and condition alignment. This results in increased generation diversity, especially at high guidance scales, with minimal loss of sample quality. This method has shown good results in the context of classifier-free guidance. However, it might also be possible to successfully incorporate it into GeoGuide.

3. Diffusion models

Diffusion models are generative algorithms that produce new samples using an iterative denoising procedure. We start from Gaussian noise x_T and gradually produce less noisy samples $x_{T-1}, x_{T-1}, \ldots, x_0$. Ultimately, x_0 comes from the data manifold. In each time step t, we have a certain noise level, and x_t is a mixture of signal x_0 and Gaussian noise ε. The time step t determines the level of noise. Diffusion models are trained using random elements and time steps to produce denoised x_{t-1} from x_t. This process is usually modeled by U-Net [4].

Diffusion models use two processes: the forward and reverse diffusion process. The first is simple. Let $q(x_0)$ denote the data distribution $x_0 \sim q(x_0)$. The forward diffusion process adds a small amount of Gaussian noise to the

sample in T steps, producing a sequence $x_0, \dots, x_T$. Such process is controlled by $\{\beta_t \in (0,1)\}_{t=1}^T$:

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

Such formula allows to obtain $x_t \sim q(x_t|x_0)$ in one step instead apply repeatedly q:

$$\begin{aligned} q(x_t|x_0) &= \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I) \\ &= \sqrt{\bar{\alpha}_t}x_0 + \varepsilon\sqrt{1-\bar{\alpha}_t},\ \varepsilon \sim \mathcal{N}(0, I), \end{aligned} \tag{2}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=0}^t \alpha_t$.

Typically we choose the schedule β_t in such a way that $\bar{\alpha}_T$ is close to 0, which implies that $q(x_T|x_0)$ is close to the distribution $\mathcal{N}(0, I)$. A simplest schedule for $\bar{\alpha}_t$ which satisfies this condition can be given by $\bar{\alpha}_t = 1 - t/T$. Additionally, we usually assume that the β_t is increasing so that we have more denoising steps at the end of the backward process, which yields better quality of generated images (while we simultaneously allow larger denoising steps at the beginning of the backward process) [4, 8].

By applying Bayes' theorem, it can be determined that the posterior distribution $q(x_{t-1}|x_t, x_0)$ is a Gaussian, characterized by the mean $\tilde{\mu}_t(x_t, x_0)$ and the variance $\tilde{\beta}_t$ as specified below:

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I),$$

where $\tilde{\mu}_t(x_t, x_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t$ and $\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$.

Theoretically, we can draw samples from the data distribution $q(x_0)$. We begin by sampling from $q(x_T)$ and then proceed by sampling the reverse steps $q(x_{t-1}|x_t)$ until we arrive at x_0. With appropriate choices for β_t and T, the distribution $q(x_T)$ approximates an isotropic Gaussian distribution, making the sampling of x_T straightforward.

Since the data distribution is unknown, we use a neural network to approximate $q(x_{t-1}|x_t)$. In the reverse diffusion process, we approximate these conditional probabilities. In [11] the authors show that $q(x_{t-1}|x_t)$ approaches a diagonal Gaussian distribution as $T \to \infty$ and, correspondingly, $\beta_t \to 0$. In the reverse diffusion process, we train a neural network to predict a mean μ_θ and a diagonal covariance matrix $\gamma_t I$:

$$p(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \gamma_t I).$$

To ensure that $p(x_0)$ captures the actual data distribution $q(x_0)$, we can optimize the corresponding variational lower bound. Such cost function is theretialy motivated, but in practice [4] propose to do not directly parameterize $\mu_\theta(x_t, t)$ as a neural network, but instead train a model $\varepsilon_\theta(x_t, t)$ to predict ε from equation (2).

The following outlines the simplified objective:

$$\mathcal{L} := \mathbb{E}_{t\sim[1,T], x_0\sim q(x_0), \varepsilon\sim\mathcal{N}(0,I)}\left[\|\varepsilon - \varepsilon_\theta(x_t, t)\|^2\right]$$

During sampling, we can use substitution to derive $\mu_\theta(x_t, t)$ from $\varepsilon_\theta(x_t, t)$:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(x_t, t)\right).$$

It is important to note that $\mathcal{L}$ does not offer any learning signal for γ_t. According to [4], rather than learning γ_t, it can be set to a constant value, either β_t or $\tilde{\beta}_t$. These constants represent the upper and lower limits for the actual reverse-step variance.

4. Classifier guidance and GeoGuide

This section presents our geometric approach to guidance in diffusion models. We start with the general idea behind the guidance of diffusion models, and then we proceed with the description of ADM-G and GeoGuide.

Let us first recall that given a controlling variance schedule $\beta_t \in (0,1)$, in the forward process, we start from the point x_0 in the data manifold M, and put

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\varepsilon_t, \text{ where } \varepsilon_t \sim \mathcal{N}(0, I).$$

Finally, we return x_T. In the backward pass, we start with randomly chosen $x_T \sim \mathcal{N}(0, I)$, and put $x_{t-1} = \mu(t, x_t) + \gamma_t\varepsilon_t$, where γ_t are constants, μ is a deep network (typically given by U-NET) and $\varepsilon_t \sim \mathcal{N}(0, I)$. The function μ and the constants γ_t are chosen so that the trajectories of the forward and backward pass constructs cannot be distinguished.

The task of guidance lies in generating data from a pre-trained diffusion model from a given class y. To do so, we usually train a classifier on $p(y|x)$ (optimally also on elements from y with added noise), and adjust the backward trajectory by the rescaled gradient of the classifier:

$$x_{t-1} = \mu(t, x_t) + \sqrt{\gamma_t}\varepsilon_t + s \cdot A(p(y|x), \nabla p(y|x)),$$

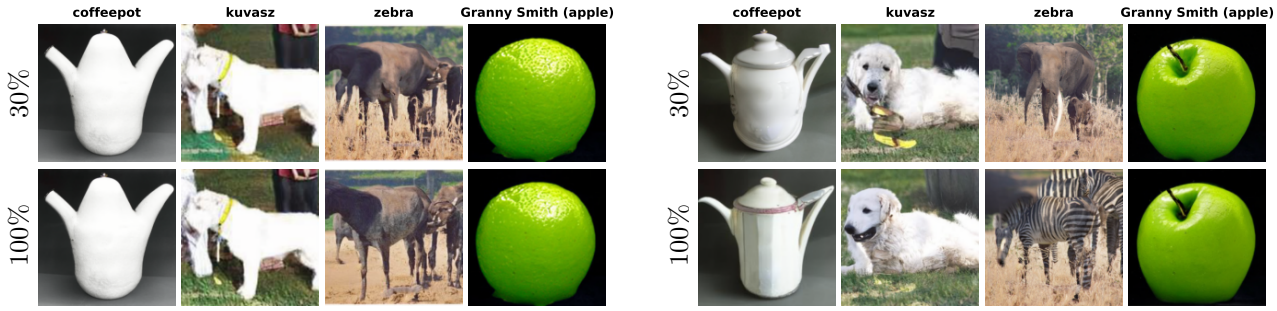
where s is the scaling parameter. The problem lies in choosing a function A that would lead to the optimal generation of points of class y. In the case of ADM-G (see Algorithm 1) we have

$$A = \gamma_t \nabla \log p(y|x) = \frac{\gamma_t}{p(y|x)} \cdot \nabla p(y|x),$$

while in the case of our model GeoGuide (see Algorithm 2) we have

$$A = \frac{\sqrt{D}}{T} \cdot \frac{\nabla p(y|x)}{\|\nabla p(y|x)\|}.$$

Before presenting the justifications for both ADM-G and GeoGuide, observe that the complexity of both adjustments is similar.



(a) ADM-G

(b) GeoGuide

Figure 3. Comparison of results when guidance is turned off after first 30% of iterations vs fully guided samples. ADM-G is not effective during last 70% of iterations, whereas GeoGuide is still significantly improving quality of generated images.

Classifier Guidance in ADM-G A key characteristic of diffusion models is their ability to generate elements based on arbitrary classes. A classifier $p(y|x)$ can enhance a diffusion generator. As demonstrated in [3, 11, 12], a pre-trained diffusion model can be conditioned using classifier gradients. Specifically, a classifier $p_\phi(y|x_t, t)$ can be trained on noisy images x_t, and the gradients $\nabla_{x_t} \log p_\phi(y|x_t, t)$ can then be used to steer the diffusion sampling process towards a specific class label y. For simplicity, we adopt the notation $p_\phi(y|x_t, t) = p_\phi(y|x_t)$ and $\varepsilon_\theta(x_t, t) = \varepsilon_\theta(x_t)$, acknowledging that these represent distinct functions for each time step t and that during training, the models must be conditioned on the time step t.

Let us assume that we have an unconditional reverse noising process $p_\theta(x_t|x_{t+1})$. To incorporate the label y into the process, we can sample each transition as follows:

$$p_{\theta,\phi}(x_t|x_{t+1}, y) = Z p_\theta(x_t|x_{t+1}) p_\phi(y|x_t),$$

where Z serves as a normalizing constant [3]. In practical applications [12], this can be approximated by a perturbed Gaussian distribution. In this section, we will revisit this derivation follow [3].

Note that our diffusion model estimates the prior time step x_t from the subsequent time step x_{t+1} utilizing a Gaussian distribution ($\Sigma = \gamma_{t+1} I$):

$$\begin{aligned} p_\theta(x_t|x_{t+1}) &= \mathcal{N}(\mu(t+1, x_{t+1}), \Sigma), \\ \log p_\theta(x_t|x_{t+1}) &= -\frac{1}{2}(x_t - \mu)^T \Sigma^{-1}(x_t - \mu) + C. \end{aligned} \tag{3}$$

It can be assumed that $\log p_\phi(y|x_t)$ exhibits low curvature compared to Σ^{-1}. This assumption holds with infinite diffusion steps, where $\|\Sigma\| \to 0$. Under these conditions, $\log\, p_\phi(y|x_t)$ can be approximated by performing a Taylor expansion around $x_t = \mu$ as:

$$\begin{aligned} &\log p_\phi(y|x_t) = \\ &\approx \log p_\phi(y|x_t)|_{x_t=\mu} + (x_t - \mu)\nabla_{x_t} \log p_\phi(y|x_t)|_{x_t=\mu} \\ &= (x_t - \mu)g + C_1, \end{aligned} \tag{4}$$

where $g = \nabla_{x_t} \log p_\phi(y|x_t)|_{x_t=\mu}$, and $C_1$is a constant. Therefore

$$\begin{aligned} &\log(p_\theta(x_t|x_{t-1}) p_\theta(y|x_t)) \\ &\approx -\frac{1}{2}(x_t - \mu)^T \Sigma^{-1}(x_t - \mu) + (x_t - \mu)g + C_2 \\ &= \log p(z) + C_4\ , z \sim \mathcal{N}(\mu + \Sigma g, \Sigma) \end{aligned} \tag{5}$$

The constant term C_4 can be disregarded, as it includes in the normalizing coefficient Z. Consequently, we have determined that the conditional transition operator can be approximated by a Gaussian, akin to the unconditional transition operator, but with its mean adjusted by Σ. Algorithm 1 outlines the related sampling algorithm. In [3], the authors introduce an optional scale factor s for the gradients.

Algorithm 1 ADM-G Classifier guided diffusion sampling, given a diffusion model $\mu(t, x_t), \gamma_t$, classifier $p(y|x_t)$, and gradient scale s.

Input: class label y, gradient scale s
$x_T \leftarrow$ sample from $N(0, I)$
for $t \leftarrow T$ to 1 **do**
  ε_t sample from $\mathcal{N}(0, I)$
  $A_t = \gamma_t \nabla \log p(y|x_t)$
  $x_{t-1} = \mu(t, x_t) + \sqrt{\gamma_t}\varepsilon_t + sA_t$
end for
return x_0

Motivation of GeoGuide This part introduces GeoGuide, which takes advantage of the metric properties of the underlying space rather than depending on probability theory.

Algorithm 2 GeoGuide: Classifier guided diffusion sampling, given a diffusion model $\mu(t, x_t), \gamma_t$, classifier $p(y|x_t)$, and gradient scale s.

Input: class label y, gradient scale s
$x_T \leftarrow$ sample from $N(0, I)$
for $t \leftarrow T$ to 1 **do**
  ε_t sample from $\mathcal{N}(0, I)$
  $A_t = \frac{\sqrt{D}}{T} \frac{\nabla p(y|x_t)}{\|\nabla p(y|x_t)\|}$
  $x_{t-1} = \mu(t, x_t) + \sqrt{\gamma_t}\varepsilon_t + sA_t$
end for
return x_0

We can interpret the forward diffusion process as a stochastic process that starts with the data manifold $M \subset \mathbb{R}^D$ and ends in the distribution $\mathcal{N}(0, I)$. In the backward process, we try to emulate the behavior of the forward process by reversing the time direction.

The forward diffusion process adds a small amount of Gaussian noise to the rescaled sample in T steps, producing a sequence $x_0, \ldots, x_T$. Using (2) we know that

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon \in \sqrt{\bar{\alpha}_t}M + \varepsilon, \text{ where } \varepsilon \sim \mathcal{N}(0, I).$$

Consequently,

$$d(x_t; \sqrt{\bar{\alpha}_t}M) \leq d(x_t, \sqrt{\bar{\alpha}_t}x_0) = \sqrt{1-\bar{\alpha}_t}\|\varepsilon\|.$$

Since in fact we only add noise to the element x_0, the closest element to $\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon$ from $\sqrt{\bar{\alpha}_t}M$ would typically be $\sqrt{\bar{\alpha}_t}x_0$. Thus, we obtain the approximation

$$d(x_t; \sqrt{\bar{\alpha}_t}M) \approx \sqrt{1-\bar{\alpha}_t}\|\varepsilon\|.$$

Since $\varepsilon \sim \mathcal{N}(0, I)$, and the dimension D of the space is large, by the law of large numbers, we obtain $\|\varepsilon\| \approx \sqrt{D}$. In conclusion, we see that the distance of the flow of the diffusion process from the (rescaled) data manifold is given by

$$d(x_t; \sqrt{\bar{\alpha}_t}M) \approx \sqrt{1-\bar{\alpha}_t} \cdot \sqrt{D}. \tag{6}$$

Since the diffusion process satisfies the metric criterion described above, our intuition suggests that any perturbations to achieve the desired changes should be implemented without changing this criterion. Thus, modifications that consider only the gradient are sub-optimal, as the value of the gradient of the classifier is inconsistent on the trajectory, see Figure 2. Consequently, we postulate that the norm of perturbation of the backward process should be consistent throughout the backward process to influence the trajectory evenly throughout the backward process.

Consequently, the main idea behind GeoGuide lies in the observation, that if during denoising we guide the model by the adjustment with norm proportionally small to the above distance, the trajectories would still satisfy (6).

Thus assume that we have a deterministic function $v(x)$ by applying which we would like to modify the trajectory of the backward process, where by the default we may think of $v(x) = \nabla p(y|x)$. Then we would normalize the norm of v to make it proportional to $\sqrt{D(1-\bar{\alpha}_t)}$, which since $\bar{\alpha}_T \approx 1$, for large t is close to $\sqrt{D}$. On the other hand, since this perturbation is deterministic, we would also normalize it by the number of steps T in the diffusion process. Finally, the adjustment will be given by

$$A_t = \frac{\sqrt{D}}{T}\sqrt{1-\bar{\alpha}_t}\frac{v(x)}{\|v(x)\|}. \tag{7}$$

Such a strategy could be easily implemented, see the Table 2, however it still could be improved by taking into account properties of the guidance given by the gradient of the classifier.

Definition of GeoGuide In the above reasoning, we have taken an arbitrary perturbation $v(x)$, which does not have to be consistent with the geometry of the data manifold M. Thus, if the perturbation (at least near M) is tangent to the manifold M, we can add a much larger perturbation and still not destroy the distance from the manifold given by (6). Observe that when we use the guidance in the backward process, we are close to the elements of the given class, and consequently, the gradient $\nabla p(y|x)$ of the classifier becomes tangent to the manifold M. Consequently, applying a larger constant than the baseline for classifier guidance does not lead to the unwanted behavior of leaving the predicted distance in (6) from the manifold. Thus, we can take a function of t which for large t is similar to the previous $\sqrt{D(1-\bar{\alpha}_t)}/T$ (as for large t we are far from the data manifold), while for small t our trajectory is close to the manifold M (the gradient of classifier becomes tangent to M) and we can choose higher values. Since $1-\bar{\alpha}_T \approx 1$, the simplest form of such strategy is given by[1]

$$A_t = \frac{\sqrt{D}}{T}\frac{\nabla p(y|x)}{\|\nabla p(y|x)\|},$$

which leads to GeoGuide, see Algorithm 2.

5. Experiments

This section presents experiments that demonstrate the efficacy of the proposed method. The evaluation follows the protocol from ADM-G [3]. Additionally, a pre-trained model derived from the provided checkpoints and the authors' recommended sampling parameters are employed. Unless otherwise specified, experiments were conducted on ImageNet 256x256 images, sampled in 250 diffusion steps.

[1] To compute the $\nabla p/\|\nabla p\|$ in a numerically stable way we use its equality to $\nabla \log p/\|\nabla \log p\|$

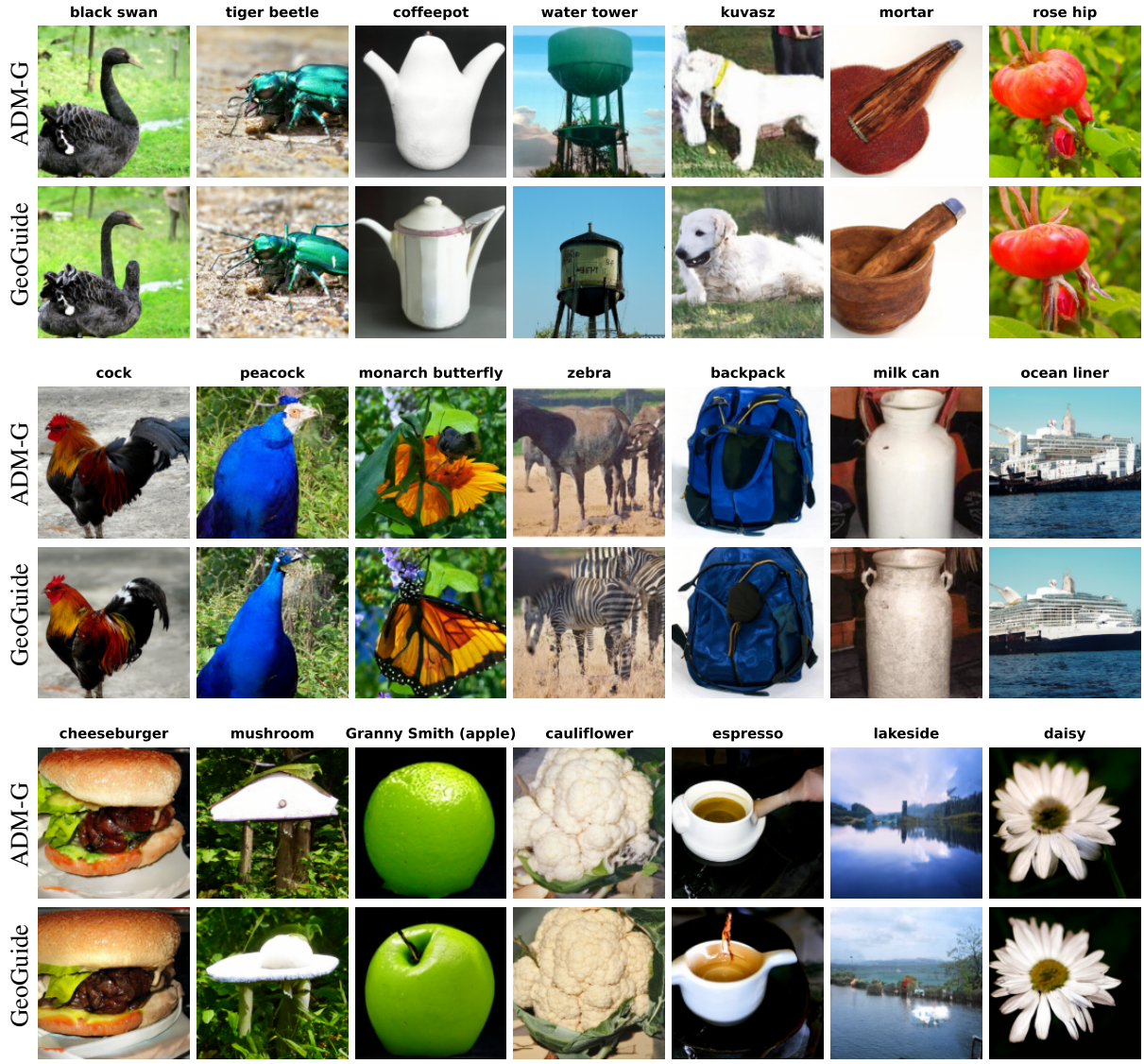


Figure 4. Images generated by guided diffusion using the same noise (random seed) and class label, with a vanilla [3] (FID 12.00, top) and a geometric (FID 7.32, bottom) guidance. Observe that images generated by GeoGuide are typically much more detailed. In our opinion, this is because the role of the classifier gradient is also important at the end of the backward process. In the ADM-G, the norm of the modification at the last steps of the process is close to zero, while in GeoGuide it stays relevant during the entire process, see Figure 2

Quantitative comparison We used the same metrics as in [3] to quantitatively evaluate our method. In Table 1 we compare the results from GeoGuide with the vanilla classifier guidance (ADM-G) [3]. To compute metrics, we generated 50000 random images and used a metric evaluation script with a reference batch provided by the authors. We are comparing models in conditional and unconditional setting. We use a classifier scales $s = 0.025$ and $s = 0.15$ respectively, which turned out to be the best values for our approach to minimize the FID score. For unconditional setting, our method shows significant improvement (7.32 vs 12.00) in terms of FID values compared to the baseline approach. In the conditional case, the improvement is much smaller, but still significant (4.06 vs 4.59). Other metrics show improvement in our method as well, if we compare cases where the classifier scale parameter of the baseline model is also optimized for FID (see Figure 7).

The classifier guide in its base form introduces a trade-off between the quality and diversity of the generated images [3], measured, for example, by FID and Recall. In our approach, this relationship is also present. Figure 6 shows how it changes depending on the varying classifier

Model	Conditional	Guidance	Scale	FID	sFID	IS	Precision	Recall
ADM [3]	✗	✗		26.21	**6.35**	39.70	0.61	0.63
ADM-G [3]	✗	✓	1.0	33.03	6.99	32.92	0.56	**0.65**
ADM-G [3]	✗	✓	10.0	12.00	10.40	95.41	0.76	0.44
GeoGuide	✗	✓	0.15	**7.32**	7.98	**243.34**	**0.77**	0.42
ADM [3]	✓	✗		10.94	6.02	100.98	0.69	**0.63**
ADM-G [3]	✓	✓	1.0	4.59	5.25	186.70	0.82	0.52
ADM-G [3]	✓	✓	10.0	9.11	10.93	**283.92**	**0.88**	0.32
GeoGuide	✓	✓	0.025	**4.06**	**5.19**	206.86	0.82	0.55

Table 1. Comparison between vanilla (ADM-G [3]) and geometric (GeoGuide) guidance. GeoGuide produce better results across metrics when compared against ADM-G variation optimized for highest FID scores. Evaluated on ImageNet 256x256 using 250 iterations during the sampling.

(a) ADM-G (b) GeoGuide (c) Dataset

Figure 5. Samples with vanilla classifier guidance [3] (FID 12.00, left) vs samples with GeoGuide (FID 7.32, middle) and samples from the training set (right). Distribution of generated samples using both guidance methods is comparable, but significantly narrower compared to samples from original dataset.

Model	Conditional	Scale	FID
GeoGuide ($\sqrt{1-\bar{\alpha}_t}$)	✗	0.15	7.47
GeoGuide	✗	0.15	7.32
GeoGuide ($\sqrt{1-\bar{\alpha}_t}$)	✓	0.025	4.78
GeoGuide	✓	0.025	4.06

Table 2. Comparison of GeoGuide with its variant given by (7), where we rescale the basic adjustment additionally by $\sqrt{1-\bar{\alpha}_t}$. Base approach achieves altogether better results. Evaluated on ImageNet 256x256 using 250 iterations during the sampling.

scale (guidance). We can observe that diversity (Recall) is the highest when we are not using guidance at all, and it is slowly decreasing as we strengthen the guidance. For the quality (FID) the relationship is reversed.

Qualitative comparison In Figure 4 we compare one-to-one samples generated with ADM-G and GeoGuide. The unconditional model trained on ImageNet 256x256 was used. Each pair was sampled using the same starting and intermediary Gaussian noises and labels. We can observe that our samples are often sharper, with more details and class-specific features.

In Figure 5 we compare the diversity of the generated images by looking at the entire batch of images of the same class. We can notice that the distribution of generated samples using both guidance methods is comparable, but significantly narrower compared to samples from the original data set.

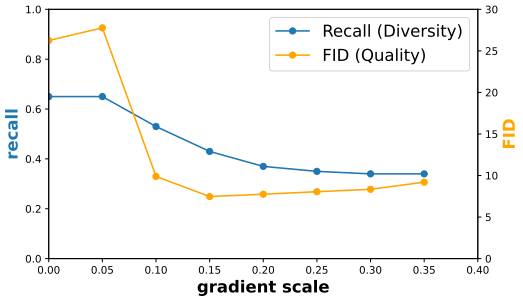


Figure 6. FID (Quality) and Recall (Diversity) trade-off in GeoGuide.

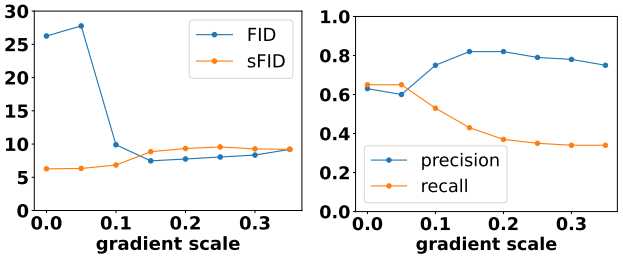


Figure 7. Change in sample quality as we vary scale of the classifier gradients for unconditional ImageNet 256×256 model. It is possible to optimize for specific metrics by modifying the scale factor accordingly.

In a handful of cases, ADM-G produces better outputs than GeoGuide. It is mostly noticeable in case of conditional model, where GeoGuide shows significantly less improvement compared to the unconditional case.

GeoGuide scaled As part of our study, we experimented with a variation of GeoGuide motivated by (7), where we take $A_t = \sqrt{D}/T\sqrt{1-\bar{\alpha}_t}\nabla p/\|\nabla p\|$ instead of the baseline $A_t = \sqrt{D}/T\nabla p/\|\nabla p\|$. We denote such model by GeoGuide($\sqrt{1-\bar{\alpha}_t}$). We thought that guiding toward a specific class is essential mainly at the early stages of the backward sampling process, when the image still forms from Gaussian noise. At later stages, it seemed like guidance should be progressively scaled down, as it would only provide irrelevant information about the class, which should already be encoded in the image itself. The numerical comparison of these two approaches is shown in Tab 2. As we can see, in the conditional and unconditional settings, the base approach is shown to produce better results.

Guidance cut-off As we observed in Figure 2 guiding factor in ADM-G quickly becomes close to 0. We think that this makes vanilla guidance effectively irrelevant for the majority of the sampling process, whereas GeoGuide can have a positive impact throughout the entire process.

Model	FID	sFID	IS	Precision	Recall
ADM-G [3]	2.97	5.09	-	0.78	0.59
Robust [5]	2.85	-	-	0.82	0.56
GeoGuide	2.83	5.17	151.63	0.80	0.61
GeoGuide + Robust	2.81	5.16	152.37	0.80	0.60

Table 3. Combining GeoGuide with robust classifier on ImageNet 128x128 using a conditional model. GeoGuide improves results independently and with a robust classifier. Missing sFID and IS values were not present in the original papers.

To observe this, we made an experiment where we turn off guidance after first 30% of the sampling iterations. As we can see in Figure 3a for ADM-G it didn't make a large difference in results, which means guidance in the following 70% is not making a large impact. In GeoGuide we can see in Figure 3b that results are much worse with guidance cut-off, so guidance stays relevant also at later iterations.

GeoGuide combined with Robust Classifier As GeoGuide can be easily incorporated into existing models, we tried to combine it with robust classifier [5], which was previously mentioned in Section 2. In Table 3 we can see that using both of these methods together further improves sampling quality and achieves the the best results in terms of quality metrics. Unfortunately, we didn't have pre-trained weights required to compare them in unconditional setting, where we would expect that results would be even more prominent.

6. Conclusion

This paper proposes GeoGuide, a method for guiding diffusion models following the distance between the denoising trajectory and the data manifold. Our metric approach can use similar guidance during the entire denoising process to obtain sharper images. In classical methods, such guidance is more critical at the beginning of the denoising process. GeoGuide is easy to execute because it depends on classifier gradient normalization and outperforms the probabilistic method ADM-G regarding FID scores and the quality of the images produced.

Acknowlegdements

The work of J. Tabor was supported by the National Centre of Science (Poland) Grant No. 2021/43/B/ST6/01456. The work of P. Spurek was supported by the National Centre of Science (Poland) Grant No. 2023/50/E/ST6/00068. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH, CI TASK, WCSS) for providing computer facilities and support within computational grant no. PLG/2024/017161.

References

[1] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023. 1

[2] Kamil Deja, Anna Kuzina, Tomasz Trzcinski, and Jakub Tomczak. On analyzing generative and denoising capabilities of diffusion-based deep generative models. *Advances in Neural Information Processing Systems*, 35:26218–26229, 2022. 2

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1, 2, 4, 5, 6, 7, 8

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[5] Bahjat Kawar, Roy Ganz, and Michael Elad. Enhancing diffusion-based image synthesis with robust classifier guidance, 2023. 2, 8

[6] Dongjun Kim, Yeongmin Kim, Se Jung Kwon, Wanmo Kang, and Il-Chul Moon. Refining generative process with discriminator guidance in score-based diffusion models, 2023. 2

[7] Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[8] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 3

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[10] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling, 2024. 2

[11] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3, 4

[12] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 4

[13] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023. 2

[14] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2