Theoretical Insights for Diffusion Guidance: A Case Study for Gaussian Mixture Models

Yuchen Wu [1] **Minshuo Chen** [2] **Zihao Li** [2] **Mengdi Wang** [2] **Yuting Wei** [1]

Abstract

Diffusion models benefit from instillation of task-specific information into the score function to steer the sample generation towards desired properties. Such information is coined as guidance. For example, in text-to-image synthesis, text input is encoded as guidance to generate semantically aligned images. Proper guidance inputs are closely tied to the performance of diffusion models. A common observation is that strong guidance promotes a tight alignment to the task-specific information, while reducing the diversity of the generated samples. In this paper, we provide the first theoretical study towards understanding the influence of guidance on diffusion models in the context of Gaussian mixture models. Under mild conditions, we prove that incorporating diffusion guidance not only boosts classification confidence but also diminishes distribution diversity, leading to a reduction in the differential entropy of the output distribution. Our analysis covers the widely adopted sampling schemes including those based on the SDE and ODE reverse processes, and leverages comparison inequalities for differential equations as well as the Fokker-Planck equation that characterizes the evolution of probability density function, which may be of independent theoretical interest.

1 Introduction

Understanding and designing algorithms for generative models that adapt to certain constraints play a crucial role in modern machine learning applications. For example, contemporary large language models — where a large model is pretrained and various natural language processing (NLP) tasks are performed based on human prompts without retraining — often demonstrate remarkable in-context learning abilities (); Text-to-image models contribute to major successes in image generators like DALL·E 2, Stable Diffusion and Imagen (Ramesh et al., 2022; Rombach et al., 2022; Saharia et al., 2022), which offer remarkable platforms for users to generate vivid images by typing in a text prompt. However, it has been observed that these models can oftentimes generate unrealistic or biased content, or not follow the users' instructions (Bommasani et al., 2021; Lučić et al., 2019; Weidinger et al., 2021). For this reason, various guided techniques have been developed to enhance the sampling qualities in accordance with users' intention (Ouyang et al., 2022; Dhariwal & Nichol, 2021; Ho & Salimans, 2022). Despite the significant empirical improvements that are observed using these guidance approaches, parameters and models are trained mainly in a trial-and-error manner. The theoretical underpinnings of these methods are still far from being mature.

1.1 Training with guidance for diffusion models

To uncover the unreasonable power of these guided approaches and better assist practice, this paper takes the first step towards this goal in the context of diffusion models. Diffusion models, which convert noise into new data instances by learning to reverse a Markov diffusion process, have become a cornerstone in contemporary generative modeling (Song et al., 2020b; Ho et al., 2020; Yang et al., 2023). Compared to alternative generative models, such as variational autoencoder or generative adversarial network, diffusion models are known to be more stable, and generate high-quality samples based on learning the gradient of the log-density function (also known as the score function). When data is multi-modal, namely, it potentially comes from multiple classes, a natural question is how to make use of these class labels for conditional synthesis. Towards this direction, (Dhariwal & Nichol, 2021) put forward the idea of *classifier guidance* — an approach to enhance the sample quality with the aid of an extra trained classifier. The classifier guidance approach combines an unconditional diffusion model's score estimate with the gradient of the log probability of a classifier. Subsequently, (Ho & Salimans, 2022) presented

[*]Equal contribution [1]Department of Statistics and Data Science, the Wharton School, University of Pennsylvania [2]Department of Electrical and Computer Engineering, Princeton University. Correspondence to: Yuchen Wu <wuyc14@wharton.upenn.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

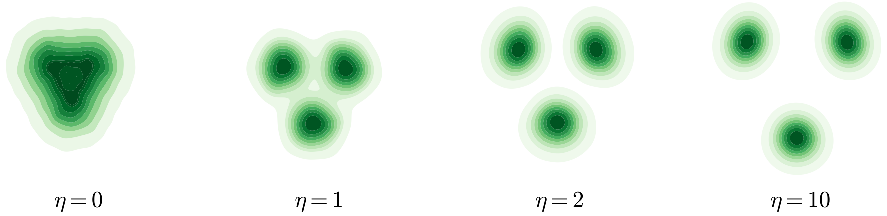

Figure 1. The effect of guidance on a three-component GMM in $\mathbb{R}^2$. Each component has weight $1/3$ and identity covariance, and the component centers are $(\sqrt{3}/2, 1/2)$, $(-\sqrt{3}/2, 1/2)$ and $(0, -1)$. The leftmost panel displays the unguided density. We increase the guidance strength from left to right. This plot imitates Figures 2 of (Ho & Salimans, 2022). Additional experiments on diverse configurations of GMMs are presented in Appendix D.

the so-called *classifier-free guidance*, which instead mixes the score estimates of an unconditional diffusion model with that of a conditional diffusion model jointly trained over the data and the label. For both guidance methods, adjusting the mixing weights of the unconditional score estimate and the other component controls the trade-off between the Fréchet Inception Distance (FID) and the Inception Score (IS) in the context of image synthesis. The resulting procedures are empirically verified to generate extremely high-fidelity samples that are at least comparable to, if not better than, other types of generative models.

One interesting feature observed for these guided procedures is an improvement in the sample quality and a decrease in the sample diversity as one increases the guidance strength (mixing weight of the other component). Specifically, (Ho & Salimans, 2022) illustrates such phenomenon numerically via a simple two-dimensional distribution comprising a mixture of three isotropic Gaussian distributions. In particular, with an increased guidance strength, the generated conditional distribution shifts its probability mass farther away from other classes, and most of the mass becomes concentrated in smaller regions, as can be seen in Figure 1. In this paper, we seek to theoretically explain this observation and provide some rigorous guarantees on how the guidance strength affects the confidence of classification and the in-class sample diversity.

1.2 Sampling from Gaussian mixture models

To allow for precise theoretical characterizations, we shall focus on the prototypical problem of sampling from Gaussian mixture models (GMMs). Specifically, we consider the data distribution p_* which takes the following form

$$p_* \overset{d}{=} \sum_{y \in \mathcal{Y}} w_y \mathsf{N}(\mu_y, \Sigma). \tag{1}$$

Here, we use y to denote the class label which takes value in a finite set $\mathcal{Y} := \{1, 2, \dots, |\mathcal{Y}|\}$. Given any class label $y \in \mathcal{Y}$, $(\mu_y, \Sigma) \in \mathbb{R}^d \times \mathbb{R}^{d \times d}$ gives the center and the covariance matrix for the Gaussian component that corresponds to y. In addition, $w_y \in \mathbb{R}_{\geq 0}$ stands for the component weight for class $y \in \mathcal{Y}$, which satisfies $\sum_{y \in \mathcal{Y}} w_y = 1$.

In this work, we investigate two widely adopted sampling methods for diffusion models, including a stochastic differential equation (SDE) based approach called the *denoising diffusion probabilistic models* (DDPMs) (Ho et al., 2020) and an ordinary differential equation (ODE) based approach called *denoising diffusion implicit models* (DDIMs) (Song et al., 2020a). An overview of these two methods under both classifier guidance and classifier-free guidance is provided in Section 2. As shall be clear momentarily, both methods involve a tuning parameter $\eta > 0$ which controls the strength of the classifier guidance (*resp.* full-model guidance) in the classifier guidance (*resp.* classifier-free guidance) approach. The overarching goal is to understand how the guidance strength affects the sample qualities, in particular, the confidence of classification and the in-class diversity.

1.3 A glimpse of main contributions

In what follows, we highlight several of our key findings.

• Consider a Gaussian mixture models with general positions. For both DDPM and DDIM samplers with diffusion guidance, we demonstrate in Section 3, that the classification confidence — which measures the posterior probability associated with the guided class given an output sample — only increases when diffusion guidance is applied. These quantitative results (Theorems 3.3 and 3.7) are further accompanied by qualitative results (Theorems 3.6 and 3.8), titrating the exact level of influence of diffusion guidance for posterior classification accuracy. These findings offer theoretical validation for employing diffusion guidance to enhance conditional sampling.

• As for the in-class diversity, in Section 4, we analyze the impact of guidance strength on the differential entropy of the resulting distribution for DDIM samplers. It turns

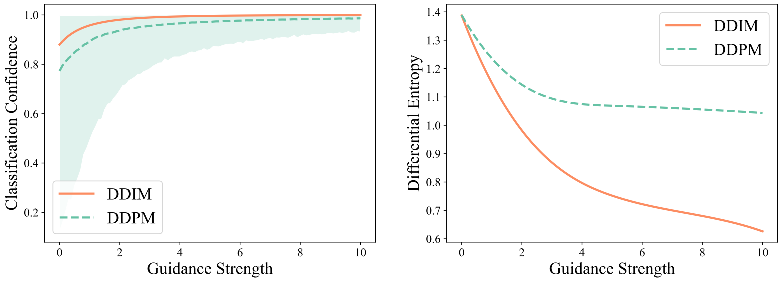

Figure 2. The effect of guidance on a symmetric GMM: $p_* = \frac{1}{2}\mathsf{N}(1,1) + \frac{1}{2}\mathsf{N}(-1,1)$. (a) In the left panel, we initiate the reverse processes at the origin, and record the classification confidence (measured by the posterior probability of class label) under different levels of guidance. For the DDPM sampler the output sample is random. We generate 10^4 samples for each guidance strength and plot the averaged classification confidence for both the DDPM and the DDIM samplers, as well as the 97.5% and 2.5% quantiles for the DDPM sampler. (b) In the right panel, we initiate the processes following a standard Gaussian distribution, and plot the differential entropy of the output distributions. For each guidance strength we also generate 10^4 samples. We adopt the function `scipy.stats.differential_entropy()` from the `scipy` module in Python to estimate the differential entropy based on these generated samples.

out that increasing the diffusion guidance always results in a reduction in the differential entropy. This offers the first theoretical explanation for the benefit of the diffusion guidance in generating more homogeneous samples.

• Finally, we exhibit that the role of the guidance strength can be complicated by an example of a three-component GMM when their means are aligned. In this case, we reveal both theoretically and numerically the existence of a phase transition in the behavior of the classification confidence as one increases the guidance strength. Cautions thus need to be exercised in practice in terms of selecting a proper guidance strength. More details can be found in Section 5.2.

Notation: For two random objects X and Y, we say $X \perp Y$ if and only if they are independent of each other. For $n \in \mathbb{N}$, we define the set $[n] = \{1, 2, \cdots, n\}$, and make the convention that $[0] = \varnothing$. We use $\sigma_{\min}(M)$ to denote the minimum eigenvalue of a matrix M.

2 Preliminaries

In this section, we introduce the basics of diffusion models, both with and without guidance. Our investigation encompasses both the SDE and ODE reverse processes. As aforementioned, there exist two primary forms of guidance, namely, the classifier guidance and the classifier-free guidance. We shall delve into a separate discussion of these two guidance forms below. As we will observe, these two forms of guidance coincide when precise access to the ground truth probability distributions is available. To enhance readers' understanding, we initiate our investigation with continuous-time processes. We later offer generalizations to discrete processes in Section 5.

2.1 Diffusion model without guidance

We begin by revisiting the concept of diffusion model without guidance. There has been a surge of recent interest and theoretical advancements to understand sampling qualities of diffusion models (e.g. Block et al. (2020); De Bortoli et al. (2021); Liu et al. (2022); De Bortoli (2022); Lee et al. (2023); Pidstrigach (2022); Chen et al. (2022b); Benton et al. (2023); Chen et al. (2022a; 2023b;a); Mei & Wu (2023); Tang & Zhao (2024); Li et al. (2023; 2024b;a)). In this paper, we focus our attention on the task of conditional sampling. More specifically, let p_* denote the data distribution over (x, y), where x is the data feature and y stands for the data label. Our goal is to sample from the conditional distribution $p_*(\cdot \mid y)$, conditioning on a label realization y. Throughout the paper, we use y to represent the label we wish to condition on. The continuous version of diffusion model consists of two processes: a forward process that converts the target distribution into noise, and a reverse process that sequentially denoises the process to reconstruct the target distribution. Throughout this paper, we set the forward process to be an Ornstein–Uhlenbeck (OU) process:

$$\mathrm{d}z_t^{\rightarrow} = -z_t^{\rightarrow}\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \qquad z_0^{\rightarrow} \sim p_*(\cdot \mid y), \tag{2}$$

where $(B_t)_{0 \le t \le T}$ is a d-dimensional standard Brownian motion. For $0 \le t \le T$, we denote by p_t the distribution of $z_t^{\rightarrow}$. The reverse process of (2) can be constructed using either an ODE or SDE implementation, which we state below:

$$\begin{aligned} \mathrm{d}z_t^{\leftarrow} &= (z_t^{\leftarrow} + \nabla \log p_{T-t}(z_t^{\leftarrow} \mid y))\mathrm{d}t, \\ \mathrm{d}\bar{z}_t^{\leftarrow} &= (\bar{z}_t^{\leftarrow} + 2\nabla \log p_{T-t}(\bar{z}_t^{\leftarrow} \mid y))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t. \end{aligned} \tag{3}$$

In the above display, $z_0^{\leftarrow}, \bar{z}_0^{\leftarrow} \sim p_{\mathrm{init}}$ for some initial distribution p_{init}, and $(B_t)_{0 \le t \le T}$ once again is the standard Brownian motion in $\mathbb{R}^d$. Hereafter, unless stated otherwise, we always take the gradient with respect to the first argument. Classical findings in probability theory (Anderson,

1982) implies that when $p_{\text{init}} = p_T(\cdot \mid y)$, it holds that $z_t^{\leftarrow} \overset{d}{=} \bar{z}_t^{\leftarrow} \overset{d}{=} z_{T-t}^{\rightarrow}$.

As a consequence, if we can implement process (3), then in principle we shall be able to generate new samples from our target distribution. To design an implementable algorithm, practioners not only apply discretization to processes in Eq. (3), but also substitute the score functions and the theoretically ideal initial distribution $p_T(\cdot \mid y)$ with their respective estimates. A standard approach for approximating $p_T(\cdot \mid y)$ is by setting $p_{\text{init}} = \mathsf{N}(0, I_d)$.

For the sake of simplicity, in the sequel we write $(z_t)_{0\le t\le T} = (z_t^{\leftarrow})_{0\le t\le T}$ and $(\bar{z}_t)_{0\le t\le T} = (\bar{z}_t^{\leftarrow})_{0\le t\le T}$ without introducing any confusion.

2.2 Classifier diffusion guidance

Classifier guidance was first proposed by (Dhariwal & Nichol, 2021) to improve the quality of images produced by diffusion models, with the aid of an extra trained classifier. To achieve this, they modify the score function to include the gradient of the logarithmic prediction probability of an auxiliary classifier. To be definite, the SDE and ODE reverse processes under classifier guidance are as follows:

$$\mathrm{d}x_t^c = \big(x_t^c + \mathcal{F}_{T-t}^c(x_t^c, y)\big)\,\mathrm{d}t, \tag{4}$$

$$\mathrm{d}\bar{x}_t^c = \big(\bar{x}_t^c + 2\mathcal{F}_{T-t}^c(x_t^c, y)\big)\,\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \tag{5}$$

where $\mathcal{F}_{T-t}^c(x_t^c, y) = s_{T-t}(x_t^c, y) + \eta\nabla_x \log c_{T-t}(x_t^c, y)$. In the above display, $\eta > 0$ is a parameter that controls the strength of the classifier guidance, $x_0^c, \bar{x}_0^c \sim p_{\text{init}}$, $s_{T-t}(x, y)$ is an estimate to $\nabla \log p_{T-t}(x \mid y)$, and $c_{T-t}(x, y)$ is a probabilistic classifier that is designed to estimate the conditional probability $p_{T-t}(y \mid x)$. When $\eta = 0$ and $s_{T-t}(x, y) = \nabla \log p_{T-t}(x \mid y)$, processes (4) and (5) reduce to their unguided counterparts.

2.3 Classifier-free diffusion guidance

Classifier guidance effectively boosts the sample quality of diffusion models. However, it requires an extra classifier, potentially introducing complexity to the model training pipeline. Classifier-free guidance is an alternative method of modifying the score functions to have the same effect as classifier guidance, but without a classifier (Ho & Salimans, 2022). To be concrete, classifier-free guidance involves the following processes:

$$\mathrm{d}x_t^f = \Big(x_t^f + \mathcal{F}_{T-t}^f(x_t^f, y)\Big)\,\mathrm{d}t, \tag{6}$$

$$\mathrm{d}\bar{x}_t^f = \Big(\bar{x}_t^f + 2\mathcal{F}_{T-t}^f(x_t^f, y)\Big)\,\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \tag{7}$$

where $\mathcal{F}_{T-t}^f(x_t^f, y) = (1+\eta)s_{T-t}(x_t^f, y) - \eta s_{T-t}(x_t^f)$, and $x_0^f, \bar{x}_0^f \sim p_{\text{init}}$. In the displayed content above, with a slight abuse of notation, we use $s_t(x)$ without the second argument to represent an estimate to the unconditional score function $\nabla_x \log p_t(x)$. Note that in situations where we have exact access to the ground truth functionals (i.e., $s_t(x, y) = \nabla_x \log p_t(x \mid y)$, $s_t(x) = \nabla_x \log p_t(x)$, and $c_t(x, y) = p_t(y \mid x)$), one can verify that $(x_t^f)_{0\le t\le T} \overset{d}{=} (x_t^c)_{0\le t\le T}$ and $(\bar{x}_t^f)_{0\le t\le T} \overset{d}{=} (\bar{x}_t^c)_{0\le t\le T}$. This result is independent of the choice of the initial distribution p_{init}. Similarly, by setting $\eta = 0$ and using the ground truth functionals, processes (6) and (7) reduce to the unguided ones.

It was observed that guidance for diffusion model, either classifier-based or classifier-free, has the effect of increasing classification confidence and decreasing sample diversity (Ho & Salimans, 2022). This paper seeks to offer a theoretical explanation of this phenomenon within the framework of GMM.

2.4 Guided diffusion for Gaussian mixture models

Under the GMM as stated in Eq (1), both the score functions and the logarithmic class probabilities admit closed-form expressions, and we shall adopt these ground truth functionals to construct our samplers. Namely, throughout this paper, we set

$$s_t(x, y) = -\Sigma_t^{-1}x + e^{-t}\Sigma_t^{-1}\mu_y, \tag{8}$$

$$\nabla_x \log c_t(x, y) = e^{-t}\Sigma_t^{-1}\mu_y - \sum_{y'\in\mathcal{Y}} e^{-t}q_t(x, y')\Sigma_t^{-1}\mu_{y'},$$

where $\Sigma_t = e^{-2t}\Sigma + (1 - e^{-2t})I_d$, and

$$q_t(x, y) := \tag{9}$$

$$\frac{w_y \exp\big(e^{-t}\langle\Sigma_t^{-1}\mu_y, x\rangle - e^{-2t}\langle\mu_y, \Sigma_t^{-1}\mu_y\rangle/2\big)}{\sum_{y'\in\mathcal{Y}} w_{y'} \exp\big(e^{-t}\langle\Sigma_t^{-1}\mu_{y'}, x\rangle - e^{-2t}\langle\mu_{y'}, \Sigma_t^{-1}\mu_{y'}\rangle/2\big)}.$$

Note that $q_t(x, y)$ is the posterior probability of having label y, upon observing $x = e^{-t}x_* + \sqrt{1 - e^{-2t}}g$, where $x_* \sim p_*$, $g \sim \mathsf{N}(0, I_d)$ and $x_* \perp g$. When the functionals listed in Eq. (8) are adopted to construct the diffusion model samplers as listed in Eq. (4)-(7), obviously we have $(x_t^f)_{0\le t\le T} \overset{d}{=} (x_t^c)_{0\le t\le T}$ and $(\bar{x}_t^f)_{0\le t\le T} \overset{d}{=} (\bar{x}_t^c)_{0\le t\le T}$.

In fact, in this case the classifier-based and the classifier-free diffusion models share the same diffusion and drift terms. Due to this observation, in the remainder of the paper we unify the notations by setting

$$(\bar{x}_t)_{0\le t\le T} = (\bar{x}_t^c)_{0\le t\le T} = (\bar{x}_t^f)_{0\le t\le T},$$

$$(x_t)_{0\le t\le T} = (x_t^c)_{0\le t\le T} = (x_t^f)_{0\le t\le T},$$

treating classifier-based and classifier-free guidance as the same algorithm.

Plugging Eq. (8) into Eq. (4)-(7), we get

$$\mathrm{d}x_t = \big(x_t - \Sigma_{T-t}^{-1}x_t + (1+\eta)e^{-T+t}\Sigma_{T-t}^{-1}\mu_y \tag{10}$$

$$- \eta e^{-T+t} \sum_{y' \in \mathcal{Y}} q_{T-t}(x_t, y') \Sigma_{T-t}^{-1} \mu_{y'} \big) \mathrm{d}t,$$

$$\mathrm{d}\bar{x}_t = \big(\bar{x}_t - 2\Sigma_{T-t}^{-1} \bar{x}_t + 2(1+\eta) e^{-T+t} \Sigma_{T-t}^{-1} \mu_y \quad (11)$$
$$- 2\eta e^{-T+t} \sum_{y' \in \mathcal{Y}} q_{T-t}(\bar{x}_t, y') \Sigma_{T-t}^{-1} \mu_{y'} \big) \mathrm{d}t + \sqrt{2} \mathrm{d}B_t.$$

3 Effect of guidance on classification confidence

As our first contribution, we offer a theoretical explanation for the phenomenon where a diffusion model with guidance directs generated samples toward a region with higher confidence, in contrast to samples generated by the unguided counterpart. To measure such confidence, we propose to examine the posterior probability

$$\mathcal{P}(x, y) := q_0(x, y) = \frac{w_y \exp\big(\langle \Sigma^{-1} \mu_y, x \rangle - \langle \mu_y, \Sigma^{-1} \mu_y \rangle / 2 \big)}{\sum_{y' \in \mathcal{Y}} w_{y'} \exp\big(\langle \Sigma^{-1} \mu_{y'}, x \rangle - \langle \mu_{y'}, \Sigma^{-1} \mu_{y'} \rangle / 2 \big)} \quad (12)$$

along the trajectory of the diffusion process as defined in Eq. (9). We show that diffusion guidance with a nonnegative guidance strength can only increase the posterior probability, given that the component centers exhibit limited correlation. Our formal assumptions are provided below.

Assumption 3.1. We impose the following conditions on model (1):

1. There exists $\mu_0 \in \mathbb{R}^d$, such that for all $y' \in \mathcal{Y}$, it holds that $|\langle \mu_y - \mu_0, \mu_{y'} - \mu_0 \rangle| \leq \varepsilon$, for some small positive constant ε. We further assume that $\varepsilon \leq \|\mu_y - \mu_0\|_2^2 / 3$.
2. Assume $w_{y'}$ is strictly positive for all $y' \in \mathcal{Y}$.
3. The GMM has an isotropic covariance: $\Sigma = I_d$.

Remark 3.2. If d is large, then the first point of Assumption 3.1 is typically satisfied when the component centers are independently generated from certain prior distribution. For instance, one can verify that the assumption is satisfied with high probability if $(\mu_{y'})_{y' \in \mathcal{Y}} \sim_{i.i.d.} \mathrm{Unif}(\mathbb{S}^{d-1})$ with a sufficiently large d, where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$.

The dynamics of $\mathcal{P}(x_t, y)$ can be represented through either an ODE or an SDE, depending on whether we utilize the ODE or the SDE framework. We explore further details in the remainder of this section.

3.1 Effect on the ODE reverse process

In this section, we analyze the impact of guidance on the ODE reverse process, as defined in Eq. (10). Our main result for this part delineates the impact of guidance on the ODE reverse process in terms of classification confidence, which we present as Theorem 3.3 below.

Theorem 3.3. *We assume model* (1) *and Assumption 3.1. Recall that x_0 and z_0 are the initializations of the ODE-based samplers as defined in Eq.* (3) *and* (10)*, respectively. In addition, we assume $\langle x_0, \mu_y - \mu_{y'} \rangle \geq \langle z_0, \mu_y - \mu_{y'} \rangle$ for all $y' \in \mathcal{Y}$*[1]*. Then for any $\eta \geq 0$ and all $t \in [0, T]$, it holds that*

$$\mathcal{P}(x_t, y) \geq \mathcal{P}(z_t, y).$$

Theorem 3.3 implies that when the processes have the same initialization, the classification confidence associated with the guided process remains no smaller than that associated with the unguided process along the entire diffusion trajectory. It therefore validates the empirical observation regarding diffusion guidance.

Our main proof idea. In order to offer some theoretical insights while at the same time maintaining brevity, we present a proof sketch of Theorem 3.3 here, and delay the majority of technical details to Appendix A.1. First, taking the inner product of the derivative given in Eq. (10) and the mean vector difference $\mu_y - \mu_{y'}$ for some $y' \in \mathcal{Y}$, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} \langle x_t, \mu_y - \mu_{y'} \rangle = e^{-(T-t)} \|\mu_y\|_2^2 - e^{-(T-t)} \langle \mu_y, \mu_{y'} \rangle$$
$$+ \eta e^{-(T-t)} (1 - q_{T-t}(x_t, y)) \|\mu_y - \mu_0\|_2^2 \quad (13)$$
$$+ \eta e^{-(T-t)} q_{T-t}(x_t, y') \|\mu_{y'} - \mu_0\|_2^2 + \mathcal{E}_t,$$

where $\mathcal{E}_t$ is a function of (x_t, t), which satisfies $|\mathcal{E}_t| \leq 3\eta e^{-(T-t)} (1 - q_{T-t}(x_t, y)) \varepsilon$. A detailed derivation of Eq. (13) is given in Appendix A. Using the assumption that $\varepsilon \leq \|\mu_y\|_2^2 / 3$, one can obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} \langle x_t, \mu_y - \mu_{y'} \rangle \geq e^{-(T-t)} \|\mu_y\|_2^2 - e^{-(T-t)} \langle \mu_y, \mu_{y'} \rangle$$
$$+ \eta e^{-(T-t)} (1 - q_{T-t}(x_t, y)) (\|\mu_y - \mu_0\|_2^2 - 3\varepsilon). \quad (14)$$

As for the unguided process $(z_t)_{0 \leq t \leq T}$, similarly we get

$$\frac{\mathrm{d}}{\mathrm{d}t} \langle z_t, \mu_y - \mu_{y'} \rangle = e^{-(T-t)} \|\mu_y\|_2^2 - e^{-(T-t)} \langle \mu_y, \mu_{y'} \rangle. \quad (15)$$

Eq. (14) and (15) motivate us to employ the ODE comparison theorem (McNabb, 1986), which we present below for readers' convenience.

Lemma 3.4 (ODE comparison theorem)**.** *Suppose $f(t, u)$ is continuous in (t, u) and Lipschitz continuous in u. Suppose $u(t)$, $v(t)$ are C^1 for $t \in [0, T]$, and satisfy*

$$u'(t) \leq f(t, u(t)), \qquad v'(t) = f(t, v(t)).$$

In addition, we assume $u(0) \leq v(0)$. Then $u(t) \leq v(t)$ for all $t \in [0, T]$.

[1] Note that this conditions is fulfilled when $x_0 = z_0$.

As a direct consequence of Lemma 3.4, Eq. (14) and (15), we derive the following lemma:

Lemma 3.5. *We adopt both model* (1) *and Assumption 3.1. We also assume $\langle z_0, \mu_y - \mu_{y'} \rangle \leq \langle x_0, \mu_y - \mu_{y'} \rangle$. Then, it holds that $\langle x_t, \mu_y - \mu_{y'} \rangle \geq \langle z_t, \mu_y - \mu_{y'} \rangle$ for all $t \in [0, T]$.*

Our proof of Theorem 3.3 makes key use of Lemma 3.5, and offers a qualitative comparison between diffusion model with guidance and the original diffusion model. We refer the readers to Appendix A.1 for a complete proof of Theorem 3.3. We also prove a result below which quantitatively measures the role of guidance. We provide the proof of Theorem 3.6 in Appendix A.2.

Theorem 3.6. *Under the assumptions of Theorem 3.3, for any $\eta \geq 0$, it holds that*

$$\begin{aligned}\mathcal{P}(x_T, y) \geq & \frac{\mathcal{P}(z_T, y)}{\mathcal{P}(z_T, y) + (1 - \mathcal{P}(z_T, y)) \cdot \exp(-\mathcal{U})} \\ \geq & \mathcal{P}(z_T, y).\end{aligned}$$

In the above display, $\mathcal{U} \in \mathbb{R}_{\geq 0}$ is any real number that satisfies

$$\begin{aligned}\mathcal{U} < & \langle x_0 - z_0, \mu_y - \mu_{y'} \rangle \\ & + (1 - e^{-T}) \cdot \eta e^{-\Delta/8} (\|\mu_y - \mu_0\|_2^2 - 3\varepsilon) \\ & \times \min \Big\{ \mathcal{F}\big(\max_{0 \leq t \leq T} \mathcal{P}(z_t, y), \mathcal{U} \big), \, \xi_w \Big\}.\end{aligned}$$

Here,

$$\begin{aligned}\mathcal{F}(p, u) &= \frac{(1-p)e^{-u}}{p + (1-p)e^{-u}}, \\ \xi_w &= 1 - w_y / (w_y + \min_{y' \neq y} w_{y'}), \\ \Delta &= \max_{y' \in \mathcal{Y}} |\|\mu_y\|_2^2 - \|\mu_{y'}\|_2^2|.\end{aligned} \tag{16}$$

Note that the lower bound above (with an optimal choice of $\mathcal{U}$) converges to 1 as $\eta \to \infty$. In addition, for a sufficiently large η it holds that

$$\mathcal{P}(x_T, y) \geq 1 - \frac{-C_0 - \mathit{logit}(\mathcal{P}(x_0, y)) + \log \eta}{\eta C_1},$$

where $C_0 = \min_{y' \in \mathcal{Y}} (1 - e^{-T}) \langle \mu_y, \mu_y - \mu_{y'} \rangle$, $C_1 = e^{-\Delta/8} (\|\mu_y - \mu_0\|_2^2 - 3\varepsilon)$, and $\mathit{logit}(p) = \log(p/(1-p))$.

The first part of Theorem 3.6 quantifies the effect of guidance strength on $\mathcal{P}(x_t, y)$ and provides lower bounds with respect to the non-guided process $\mathcal{P}(z_t, y)$ everywhere along the diffusion path. We note that this lower bound serves as an initial attempt and might be still far from tight. We leave the improvement to future works. The second part of Theorem 3.6 implies that $\mathcal{P}(x_T, y) \to 1$ as $\eta \to \infty$, and the convergence rate is at least $1 - O(\eta^{-1} \log \eta)$. In another word, if the guidance strength is chosen to be very large, then the classification confidence will be close to one.

3.2 Effect on the SDE reverse process

We then switch to consider the SDE reverse process, and we compare in this section $\mathcal{P}(\bar{x}_t, y)$ and $\mathcal{P}(\bar{z}_t, y)$, where we recall that $\{\bar{x}_t\}_{0 \leq t \leq T}$ and $\{\bar{z}_t\}_{0 \leq t \leq T}$ are defined respectively in Eq. (11) and (3). A notable distinction with the ODE result arises in the need for an SDE comparison theorem, which we state as Lemma A.1 in the appendix. Lemma A.1 enables us to establish the following theorem, the proof of which can be found in Appendix A.3.

Theorem 3.7. *We assume the assumptions of Theorem 3.3, then for any $\eta \geq 0$, almost surely we have $\mathcal{P}(\bar{x}_t, y) \geq \mathcal{P}(\bar{z}_t, y)$ for all $t \in [0, T]$.*

We also develop a quantitative comparison, presented as Theorem 3.8 below, the proof of which is deferred to Appendix A.4.

Theorem 3.8. *We assume the conditions of Theorem 3.3. Then, for any $\eta \geq 0$, almost surely we have*

$$\begin{aligned}\mathcal{P}(\bar{x}_T, y) \geq & \frac{\mathcal{P}(\bar{z}_T, y)}{\mathcal{P}(\bar{z}_T, y) + (1 - \mathcal{P}(\bar{z}_T, y)) \cdot \exp(-\bar{\mathcal{U}})} \\ \geq & \mathcal{P}(\bar{z}_T, y),\end{aligned}$$

where $\bar{\mathcal{U}}$ is any non-negative number that satisfies

$$\begin{aligned}&\bar{\mathcal{U}} < e^{-T} \langle \bar{x}_0 - \bar{z}_0, \mu_y - \mu_{y'} \rangle \\ &+ \eta (1 - e^{-2T}) e^{-\Delta/8} \min \big\{ \mathcal{F}\big(\max_{0 \leq t \leq T} \mathcal{P}(\bar{z}_t, y), \, e^T \bar{\mathcal{U}} \big), \, \xi_w \big\} \\ &\times (\|\mu_y - \mu_0\|_2^2 - 3\varepsilon),\end{aligned}$$

where we recall that $(\mathcal{F}, \xi_w, \Delta)$ are defined in Eq. (16). One can verify that the above lower bound (with an optimal choice of $\bar{\mathcal{U}}$) approaches 1 as η tends to infinity. If we fix the path initialization and the Brownian motion realization and only set $\eta \to \infty$, then the convergence rate is at least $1 - O(\eta^{-e^{-T}} (\log \eta)^{2e^{-T}})$.

Theorems 3.7 and 3.8 are counterparts of the results for the ODE reverse process that we have established in Section 3.1, indicating that adding guidance only increases the classification confidence for the SDE reverse process. Due to the stochastic nature of the SDE reverse process, the results in this section only hold almost surely.

3.3 Special case: GMM with two clusters

The results presented in Sections 3.1 and 3.2 are derived based on Assumption 3.1. It turns out that we can further relax our assumptions when the number of Gaussian components is two (i.e., $|\mathcal{Y}| = 2$), which we report in this section.

Without loss, we let $\mathcal{Y} = \{1, 2\}$, and assume guidance is towards the cluster that has label 1. Correspondingly, the GMM considered here admits the following representation:

$$w_1 \mathsf{N}(\mu_1, I_d) + w_2 \mathsf{N}(\mu_2, I_d),$$

where $w_1, w_2 \in \mathbb{R}_{\geq 0}$ satisfies $w_1 + w_2 = 1$.

To summarize, in order to establish a similar set of results for the two-component GMM, we only require the second and the third points of Assumption 3.1. We collect results for the ODE and the SDE reverse processes separately below as Theorems 3.9 and 3.10. We prove them in Appendices A.5 and A.6, respectively.

Theorem 3.9. *We assume $|\mathcal{Y}| = 2$, as well as the second and the third points of Assumption 3.1. Then the following statements regarding the ODE reverse process are true:*

1. *If $\langle x_0, \mu_1 - \mu_2 \rangle \geq \langle z_0, \mu_1 - \mu_2 \rangle$, then $\mathcal{P}(x_t, 1) \geq \mathcal{P}(z_t, 1)$ for all $t \in [0, T]$.*
2. *If $\langle x_0, \mu_1 - \mu_2 \rangle \geq \langle z_0, \mu_1 - \mu_2 \rangle$, then*

$$\begin{aligned}\mathcal{P}(x_T, 1) \geq & \frac{\mathcal{P}(z_T, 1)}{\mathcal{P}(z_T, 1) + (1 - \mathcal{P}(z_T, 1)) \cdot \exp(-\mathcal{U})} \\ \geq & \mathcal{P}(z_T, 1),\end{aligned}$$

where $\mathcal{U}$ is any non-negative number that satisfies

$$\begin{aligned}\mathcal{U} < & 2\langle x_0 - z_0, \mu \rangle + 4\eta e^{-\Delta_1/8} \|\mu\|_2^2 (1 - e^{-T}) \\ & \times \min\left\{\mathcal{F}\left(\max_{0 \leq t \leq T} \mathcal{P}(z_t, 1), \mathcal{U}\right), w_2\right\}.\end{aligned}$$

In the above display, $\mathcal{F}(\cdot)$ is defined in Eq. (16), and $\Delta_1 = |\|\mu_1\|_2^2 - \|\mu_2\|_2^2|$. The lower bound above approaches one as $\eta \to \infty$. Furthermore, the convergence rate is at least $1 - O(\eta^{-1}(\log \eta)^2)$.

Theorem 3.10. *We assume the conditions of Theorem 3.9, and consider the SDE reverse process. Then the following statements hold almost surely:*

1. *If $\langle \bar{x}_0, \mu_1 - \mu_2 \rangle \geq \langle \bar{z}_0, \mu_1 - \mu_2 \rangle$, then $\mathcal{P}(\bar{x}_t, 1) \geq \mathcal{P}(\bar{z}_t, 1)$ for all $t \in [0, T]$.*
2. *If $\langle \bar{x}_0, \mu_1 - \mu_2 \rangle \geq \langle \bar{z}_0, \mu_1 - \mu_2 \rangle$, then for all $t \in [0, T]$*

$$\begin{aligned}\mathcal{P}(\bar{x}_T, 1) \geq & \frac{\mathcal{P}(\bar{z}_T, 1)}{\mathcal{P}(\bar{z}_T, 1) + (1 - \mathcal{P}(\bar{z}_T, 1)) \cdot \exp(-\bar{\mathcal{U}})} \\ \geq & \mathcal{P}(\bar{z}_T, 1),\end{aligned}$$

where $\bar{\mathcal{U}}$ is any non-negative number such that

$$\begin{aligned}\bar{\mathcal{U}} < & 2e^{-T}\langle \bar{x}_0 - \bar{z}_0, \mu \rangle + 4\eta e^{-\Delta_1/8} \|\mu\|_2^2 (1 - e^{-2T}) \\ & \times \min\left\{\mathcal{F}\left(\max_{0 \leq t \leq T} \mathcal{P}(\bar{z}_t, 1), e^T \bar{\mathcal{U}}\right), w_2\right\},\end{aligned}$$

where we recall that $\mathcal{F}(\cdot)$ is defined in Eq. (16), and $\Delta_1 = |\|\mu_1\|_2^2 - \|\mu_2\|_2^2|$. The lower bound in the theorem converges to 1 as $\eta \to \infty$. Furthermore, the convergence rate is at least $1 - O(\eta^{-e^{-T}} (\log \eta)^{2e^{-T}})$.

The results above confirm that diffusion model with guidance always promotes classification confidence in two-component GMMs. It is interesting to note that augmenting a center component to the two-component GMM leads to complicated consequence in terms of guidance; see details in Section 5.2.

4 Effect of guidance on distribution diversity

In this section, we investigate the impact of guidance on distribution diversity. We propose to employ the *differential entropy* of probability distributions to measure diversity (Shannon, 1948). This section exclusively concentrates on the ODE reverse process. To define differential entropy, we denote by $Q(t, x)$ the probability density function of x_t, where we recall that $(x_t)_{0 \leq t \leq T}$ is defined in Eq. (10). For comparison, we also denote by $Q_0(t, x)$ the probability density function of the unguided process $(z_t)_{0 \leq t \leq T}$ defined in Eq. (3). We shall prove in appendix that the probability density functions exist for all $t \in [0, T]$ if we assume it exists at $t = 0$. Our objective is to delineate the influence of diffusion guidance on the entropy functionals, as defined below:

$$\begin{aligned}H(t) &:= -\int Q(t, x) \log Q(t, x) \mathrm{d}x, \quad 0 \leq t \leq T, \\ H_0(t) &:= -\int Q_0(t, x) \log Q_0(t, x) \mathrm{d}x, \quad 0 \leq t \leq T.\end{aligned} \tag{17}$$

Intuitively, a high entropy indicates that the distribution is spread in the space, while on the contrary, a low entropy is oftentimes associated with relatively concentrated distributions.

We propose to analyze the evolution of the entropy using the Fokker-Planck equation (Fokker, 1914), which characterizes the distributional evolution of the ODE reverse process. Readers may refer to Lemma 4.1 for a detailed exposure.

Lemma 4.1 (Fokker–Planck equation)**.** *Consider the d-dimensional SDE*

$$\mathrm{d}X_t = \mu(t, X_t)\mathrm{d}t + \sigma(t, X_t)\mathrm{d}B_t,$$

where $\mu, \sigma : \mathbb{R}_{\geq 0} \times \mathbb{R}^d \mapsto \mathbb{R}^d$ satisfies $\|\mu(t, x) - \mu(t, y)\|_2 + \|\sigma(t, x) - \sigma(t, y)\|_2 \leq C\|x - y\|_2$ for some constant C and all $x, y \in \mathbb{R}^d$. Assume that the probability density function (w.r.t. the Lebesgue measure) of X_t exists for all $t \in [0, T]$, and denote by $p(t, x)$ the probability density function for X_t. We also assume all the relevant functions are continuously differentiable, then

$$\begin{aligned}\frac{\partial}{\partial t} p(t, x) = & -\sum_{i=1}^{d} \frac{\partial}{\partial x_i} \left[\mu_i(t, x) p(t, x)\right] \\ & + \sum_{i=1}^{d} \sum_{j=1}^{d} \frac{\partial^2}{\partial x_i \partial x_j} \left[D_{ij}(t, x) p(t, x)\right],\end{aligned}$$

where $D(t,x) = \sigma(t,x)\sigma(t,x)^\top/2$.

Our theorem is stated below. A heuristic derivation based on the Fokker-Planck equation is in Appendix B.1, and a formal proof of the theorem is postponed to Appendix B.3.

Theorem 4.2. *We assume that both x_0 and z_0 have probability density functions with respect to the Lebesgue measure, and the corresponding differential entropies exist and are finite, satisfying $H_0(0) \geq H(0)$. We also assume model (1), Σ is non-degenerate, as well as the second point of Assumption 3.1. Then for all $0 \leq t \leq T$, it holds that $H_0(t) \geq H(t)$.*

In contrast to the results presented in Section 3, Theorem 4.2 does not require an isotropic covariance matrix, and places no assumptions on the component centers. Mild regularity condition is imposed on the process initialization to ensure the existence of the differential entropy.

Setting $t = T$, Theorem 4.2 says that the generated distribution under diffusion guidance has lower entropy compared to that without guidance. This corroborates the common observation displayed in Figure 1: diffusion guidance reduces diversity of the generated samples.

5 Effect of guidance on discretized process

In practice, it is essential to employ discretization to approximate the continuous-time processes. The widely adopted exponential integrator sampling scheme takes the following algorithmic implementations for DDIM (process (10)) and DDPM (process (11)), respectively:

$$\begin{aligned} X_{k+1} &= e^{\delta_k} X_k + (e^{\delta_k} - 1)\big(\nabla_x \log p_{T-t_k}(X_k, y) \\ &\quad + \eta \cdot \nabla_x \log p_{T-t_k}(y \mid X_k)\big), \qquad (18) \\ \bar{X}_{k+1} &= e^{\delta_k} \bar{X}_k + (e^{\delta_k} - 1)\big(\bar{X}_k + 2\nabla_x \log p_{T-t_k}(\bar{X}_k, y) \\ &\quad + 2\eta \nabla_x \log p_{T-t_k}(y \mid \bar{X}_k)\big) + \sqrt{2e^{\delta_k} - 2}\, W_k. \qquad (19) \end{aligned}$$

In the display above, Eq. (18) corresponds to the DDIM and Eq. (19) corresponds to the DDPM, $W_k \sim \mathsf{N}(0, I_d)$ and is independent of the previous iterates, $0 = t_0 < t_1 < \cdots < t_K \leq T$, $\delta_k > 0$ and $t_{k+1} = \sum_{i=0}^{k} \delta_i$ for all $k = 0, 1, \cdots, K-1$.

Analogously, to set up comparison, we also consider the discretized processes without guidance:

$$\begin{aligned} Z_{k+1} &= e^{\delta_k} Z_k + (e^{\delta_k} - 1)\left(Z_k + \nabla_x \log p_{T-t_k}(Z_k, y)\right), \\ \bar{Z}_{k+1} &= e^{\delta_k} \bar{Z}_k + (e^{\delta_k} - 1)\left(\bar{Z}_k + 2\nabla_x \log p_{T-t_k}(\bar{Z}_k, y)\right) \\ &\quad + \sqrt{2e^{\delta_k} - 2}\, W_k. \end{aligned}$$

We unify the discretization schemes for both the guided and the unguided processes to facilitate meaningful comparisons. In the current regime, we are able to establish results related to classification confidence and distribution diversity, which we collect below.

5.1 Results for the DDIM sampler

We first investigate the prediction confidence, and establish the following theorem. We postpone the proof of the theorem to Appendix C.1.

Theorem 5.1. *We assume model (1) and Assumption 3.1. We also assume $\langle X_0, \mu_y - \mu_{y'} \rangle \geq \langle Z_0, \mu_y - \mu_{y'} \rangle$ for all $y' \in \mathcal{Y}$. Then the following statements are true:*

1. *For all $k \in \{0\} \cup [K]$, it holds that $\mathcal{P}(X_k, y) \geq \mathcal{P}(Z_k, y)$.*
2. *We let $\Delta_{\max} = \max_{j \in \{0\} \cup [K-1]} \delta_j$, then for any $\eta \geq 0$, it holds that*

$$\begin{aligned} \mathcal{P}(X_K, y) \geq & \frac{\mathcal{P}(Z_K, y)}{\mathcal{P}(Z_K, y) + (1 - \mathcal{P}(Z_K, y)) \cdot \exp(-\mathcal{U})} \\ \geq & \mathcal{P}(Z_K, y), \end{aligned}$$

where $\mathcal{U} > 0$ is any number that satisfies

$$\begin{aligned} & \mathcal{U} - \langle X_0 - Z_0, \mu_y - \mu_{y'} \rangle \\ & < (e^{-T+t_K} - e^{-T})\Big(\eta e^{-\Delta/8}(\|\mu_y - \mu_0\|_2^2 - 3\varepsilon) \\ & \quad \times \min\{\mathcal{F}(\max_{0 \leq k \leq K} \mathcal{P}(Z_k, y), \mathcal{U}), \xi_w\}\Big), \end{aligned}$$

where we recall that $(\mathcal{F}, \xi_w, \Delta)$ are defined in Eq. (16). Furthermore, as $\eta \to \infty$, we have $\mathcal{P}(X_K, y) \geq 1 - O(\eta^{-1}(\log \eta)^2)$.

As can be seen, discretization preserves the boost of the prediction confidence under guidance. Yet we note that the discretization step size δ_k interacts with the increment of the prediction confidence: Large step size leads to a marginal increase, as demonstrated by the second item of Theorem 5.1. We can establish analogous results for the discretized DDPM sampler, which is provided in Appendix C.3.

We can also establish results on differential entropy for the DDIM sampler with discretization. For $k \in \{0, 1, \cdots, K\}$ we denote by $\mathcal{H}(k)$ the differential entropy of X_k[2] and denote by $\mathcal{H}_0(k)$ that of Z_k. Our main theorem for this part shows that under mild regularity conditions, it holds that $\mathcal{H}(k) \leq \mathcal{H}_0(k)$ for all $k \in \{0, 1, \cdots, K\}$. The proof of Theorem 5.2 is postponed to Appendix C.2.

Theorem 5.2. *We assume both X_0 and Z_0 have probability density functions with respect to the Lebesgue measure,*

[2] Namely, $\mathcal{H}(k) = -\int p_k(x) \log p_k(x) \mathrm{d}x$, where $p_k(\cdot)$ is the density function of X_k. We shall prove in Lemma C.1 that with mild assumptions this differential entropy exists.

and the corresponding differential entropies exist and are finite, satisfying $\mathcal{H}(0) \le \mathcal{H}_0(0)$. We also assume model (1)*, the second point of Assumption* 3.1*, and that Σ is non-degenerate. In addition, for all $k \in \{0\} \cup [K-1]$, we require the step sizes are small enough such that*

$$e^{\delta_k} > \frac{e^{\delta_k}-1}{\sigma_{\min}(\Sigma)\wedge 1} + \frac{(e^{\delta_k}-1)\eta \sup_{y'\in\mathcal{Y}}\|\mu_{y'}\|_2^2}{\sigma_{\min}(\Sigma)^2\wedge 1},$$

$$e^{\delta_k}-1+\frac{e^{\delta_k}-1}{\sigma_{\min}(\Sigma)\wedge 1} + \frac{(e^{\delta_k}-1)\eta \sup_{y'\in\mathcal{Y}}\|\mu_{y'}\|_2^2}{\sigma_{\min}(\Sigma)^2\wedge 1} < 1/2,$$

where $\sigma_{\min}(\Sigma)$ is the minimum eigenvalue of Σ. Then, for all $k \in \{0\} \cup [K]$ we have $\mathcal{H}(k) \le \mathcal{H}_0(k)$.

Theorem 5.2 resembles the conclusion of Theorem 4.2 by requiring relatively small step sizes. While the range of the required step size is rather technical, we anticipate the discretization error interacts with the differential entropy of the generated distribution. Indeed, in the discretized DDIM samplers, there is also complicated interplay between the strength of the guidance with the discretization step size.

5.2 A negative example of strong guidance

We demonstrate a negative impact of strong guidance under discretization. We focus on a three-component aligned Gaussian mixture model:

$$\mu_{\text{neg}} \overset{d}{=} \frac{1}{3}\mathsf{N}(-\mu, I_d) + \frac{1}{3}\mathsf{N}(0, I_d) + \frac{1}{3}\mathsf{N}(\mu, I_d), \qquad (20)$$

where $\mu \neq \vec{0}$ is the mean vector. Note that the first item of Assumption 3.1 does not hold here. We study the generation of samples corresponding to the center component $\mathsf{N}(0, I)$. For simplicity, we focus on the discretized DDIM backward process (18). The following lemma establishes the divergence of $\langle X_k, \mu\rangle$ under strong guidance.

Proposition 5.3. *Consider the Gaussian mixture model in* (20)*. There exist constants η_0 and η_0' that depend on the discretization step size, such that for any k verifying $e^{-T+t_k} \ge 1/2$,*

- *when $\eta \le \eta_0$, $|\langle X_{k+1}, \mu\rangle| < |\langle X_k, \mu\rangle|$ for $\langle X_k, \mu\rangle \neq 0$;*
- *when $\eta \ge \eta_0'$,*

$$|\langle X_{k+1}, \mu\rangle| > |\langle X_k, \mu\rangle| \quad \text{if} \quad |\langle X_k, \mu\rangle| \in (0, a];$$
$$|\langle X_{k+1}, \mu\rangle| < |\langle X_k, \mu\rangle| \quad \text{if} \quad |\langle X_k, \mu\rangle| > b,$$

where $a, b > 0$ are constants dependent on η.

The proof is deferred to Appendix C.5. A large $|\langle X_k, \mu\rangle|$ indicates a strong likelihood that X_k will be classified into one of the components $\mathsf{N}(\pm\mu, I_d)$. Therefore, Proposition 5.3 implies that there exists a phase shift on the generation of the center component. With weak guidance, the center component is condensed. However, under strong guidance, the center component tends to vanish as the generated samples are pushed towards side centers. This phenomenon entangles with the discretization step size and is demonstrated in Figure 4. As can be seen, using large guidance and discretization step sizes, the center component is split into two symmetric components. We also note that larger guidance strength is required as the discretization step size decreases for the split phenomenon to occur.

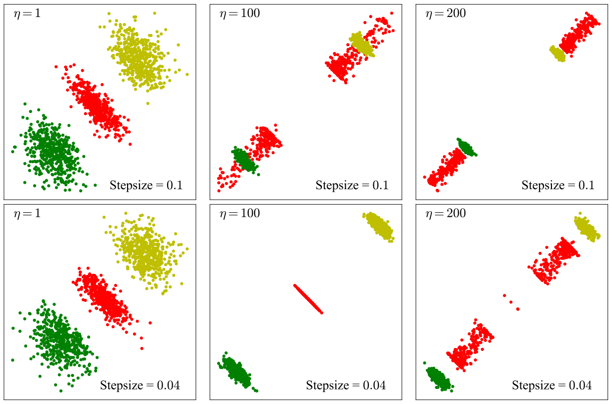

Figure 3. Negative effect of large guidance: We choose $\mu = [2, 2]^\top$ in μ_{neg}. The upper row uses a large discretization step size. Under strong guidance, the center component splits into two clusters early. The bottom row uses a small discretization step size; the center component splits under large guidance.

6 Conclusion and Discussion

In this paper, we establish the theoretical foundation for diffusion guidance in the context of sampling from Gaussian mixture models with shared covariance matrices. Under a set of mild regularity conditions, we show that guidance increases the prediction confidence along every realized path, while decreasing the overall distribution diversity. Our analysis is based on ODE and SDE comparison theorems, along with the Fokker-Planck equation that depicts the evolution of probability density functions.

We list here several interesting future directions that deserve further investigation. First, the quantitative lower bounds we present in the paper might not be tight, and a more careful examination of the guidance effect is worthy of future studies. Secondly, due to technical reasons, we currently lack a characterization of the reduction in diversity that arises from guidance for the SDE-based sampler. It is of great interest to derive similar guarantees for the SDE-based sampler. Finally, we expect our framework to go beyond sampling from GMMs and we leave this extension to future work.

Impact statement

Our research has various potential societal impacts, particularly in enhancing fairness within text-to-image synthesis using diffusion models. In practice, generative models may inherit biases from the dataset, enhancing stereotypes and limiting opportunities for minority groups (Seshadri et al., 2023). Recent efforts such as Fair Diffusion (Friedrich et al., 2023) propose fairness guidance, extending classifier-free guidance by adding an additional term representing fairness in the generation process. By tailoring guidance strength with specific prompts on fairness, such methods result in more diverse outputs. Our study provides a theoretical understanding of the impact of guidance strength and suggests caution in proper choice of guidance in fairness-oriented tasks.

Acknowledgement

Mengdi Wang is partially supported by NSF IIS-2107304, NSF CPS-2312093, ONR 1006977 and Genmab. Yuting Wei is supported in part by the NSF grants DMS-2147546/2015447, CAREER award DMS-2143215, CCF-2106778, and the Google Research Scholar Award.

References

Anderson, B. D. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Benton, J., De Bortoli, V., Doucet, A., and Deligiannidis, G. Linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*, 2023.

Block, A., Mroueh, Y., and Rakhlin, A. Generative modeling with denoising auto-encoders and langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Borel, É. *Leçons sur la théorie des fonctions*. Gauthier-Villars et fils, 1928.

Chen, H., Lee, H., and Lu, J. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. *arXiv preprint arXiv:2211.01916*, 2022a.

Chen, M., Huang, K., Zhao, T., and Wang, M. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. *arXiv preprint arXiv:2302.07194*, 2023a.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. R. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022b.

Chen, S., Daras, G., and Dimakis, A. G. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for DDIM-type samplers. *arXiv preprint arXiv:2303.03384*, 2023b.

De Bortoli, V. Convergence of denoising diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2208.05314*, 2022.

De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Fokker, A. D. Die mittlere energie rotierender elektrischer dipole im strahlungsfeld. *Annalen der Physik*, 348(5):810–820, 1914.

Friedrich, F., Schramowski, P., Brack, M., Struppek, L., Hintersdorf, D., Luccioni, S., and Kersting, K. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. Improved precision and recall metric for assessing generative models, 2019.

Lee, H., Lu, J., and Tan, Y. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pp. 946–985, 2023.

Li, G., Wei, Y., Chen, Y., and Chi, Y. Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv preprint arXiv:2306.09251*, 2023.

Li, G., Huang, Y., Efimov, T., Wei, Y., Chi, Y., and Chen, Y. Accelerating convergence of score-based diffusion models, provably. 2024a.

Li, G., Huang, Z., and Wei, Y. Towards a mathematical theory for consistency training in diffusion models. *arXiv preprint arXiv:2402.07802*, 2024b.

Liu, X., Wu, L., Ye, M., and Liu, Q. Let us build bridges: Understanding and extending diffusion generative models. *arXiv preprint arXiv:2208.14699*, 2022.

Lučić, M., Tschannen, M., Ritter, M., Zhai, X., Bachem, O., and Gelly, S. High-fidelity image generation with fewer labels. In *International conference on machine learning*, pp. 4183–4192. PMLR, 2019.

McNabb, A. Comparison theorems for differential equations. *Journal of mathematical analysis and applications*, 119(1-2):417–428, 1986.

Mei, S. and Wu, Y. Deep networks as denoising algorithms: Sample-efficient learning of diffusion models in high-dimensional graphical models. *arXiv preprint arXiv:2309.11420*, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Pidstrigach, J. Score-based generative models detect manifolds. *arXiv preprint arXiv:2206.01018*, 2022.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Rudin, W. et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1976.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.

Seshadri, P., Singh, S., and Elazar, Y. The bias amplification paradox in text-to-image generation, 2023.

Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Tang, W. and Zhao, H. Contractive diffusion probabilistic models. *arXiv preprint arXiv:2401.13115*, 2024.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., and Yang, M.-H. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

Zhu, X. On the comparison theorem for multidimensional sdes with jumps. *arXiv preprint arXiv:1006.1454*, 2010.

A Proofs related to confidence enhancement

This section contains proofs pertinent to results on guidance improving prediction confidence. We first prove Eq. (13). Note that

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\langle x_t, \mu_y - \mu_{y'}\rangle \\
=& e^{-(T-t)}\|\mu_y\|_2^2 + \eta e^{-(T-t)}\|\mu_y - \mu_0\|_2^2 - \eta e^{-(T-t)} \sum_{y''\in\mathcal{Y}} q_{T-t}(x_t, y'')\langle \mu_y - \mu_0, \mu_{y''} - \mu_0\rangle \\
&- e^{-(T-t)}\langle \mu_y, \mu_{y'}\rangle - \eta e^{-(T-t)}\langle \mu_y - \mu_0, \mu_{y'} - \mu_0\rangle + \eta e^{-(T-t)} \sum_{y''\in\mathcal{Y}} q_{T-t}(x_t, y'')\langle \mu_{y'} - \mu_0, \mu_{y''} - \mu_0\rangle \\
=& e^{-(T-t)}\|\mu_y\|_2^2 - e^{-(T-t)}\langle \mu_y, \mu_{y'}\rangle + \eta e^{-(T-t)}(1 - q_{T-t}(x_t, y))\|\mu_y - \mu_0\|_2^2 \\
&+ \eta e^{-(T-t)} q_{T-t}(x_t, y')\|\mu_{y'} - \mu_0\|_2^2 + \mathcal{E}_t,
\end{aligned}
\tag{21}
$$

where

$$
\begin{aligned}
\mathcal{E}_t =& - \eta e^{-(T-t)} \sum_{y''\in\mathcal{Y}\backslash\{y\}} q_{T-t}(x_t, y'')\langle \mu_y - \mu_0, \mu_{y''} - \mu_0\rangle - \eta e^{-(T-t)}(1 - q_{T-t}(x_t, y))\langle \mu_y - \mu_0, \mu_{y'} - \mu_0\rangle \\
&+ \eta e^{-(T-t)} \sum_{y''\in\mathcal{Y}\backslash\{y, y'\}} q_{T-t}(x_t, y'')\langle \mu_{y'} - \mu_0, \mu_{y''} - \mu_0\rangle.
\end{aligned}
$$

By triangle inequality and Assumption 3.1 it holds that $|\mathcal{E}_t| \le 3\eta e^{-(T-t)}(1 - q_{T-t}(x_t, y))\varepsilon$. This completes the proof of Eq. (13).

A.1 Proof of Theorem 3.3

Observe that

$$
\begin{aligned}
\mathcal{P}(x_t, y) &= \frac{w_y}{w_y + \sum_{y'\in\mathcal{Y}, y'\neq y} w_{y'} \exp\left(\langle x_t, \mu_{y'} - \mu_y\rangle - \|\mu_{y'}\|_2^2/2 + \|\mu_y\|_2^2/2\right)}, \\
\mathcal{P}(z_t, y) &= \frac{w_y}{w_y + \sum_{y'\in\mathcal{Y}, y'\neq y} w_{y'} \exp\left(\langle z_t, \mu_{y'} - \mu_y\rangle - \|\mu_{y'}\|_2^2/2 + \|\mu_y\|_2^2/2\right)}.
\end{aligned}
$$

According to Lemma 3.5, we have $\langle x_t, \mu_y - \mu_{y'}\rangle \ge \langle z_t, \mu_y - \mu_{y'}\rangle$ for all $y' \in \mathcal{Y}\backslash\{y\}$, hence $\mathcal{P}(x_t, y) \ge \mathcal{P}(z_t, y)$ for all $t \in [0, T]$. This completes the proof of Theorem 3.3.

A.2 Proof of Theorem 3.6

PROOF OF THE FIRST RESULT

The idea is to first establish an upper bound for $q_{T-t}(x_t, y)$, then in turn use it to lower bound the effect of guidance. Notice that

$$
\begin{aligned}
q_{T-t}(x_t, y) =& \frac{w_y}{w_y + \sum_{y'\neq y} w_{y'} \exp\left(e^{-(T-t)}\langle x_t, \mu_{y'} - \mu_y\rangle - e^{-2(T-t)}(\|\mu_{y'}\|_2^2 - \|\mu_y\|_2^2)/2\right)} \\
=& \frac{\widetilde{q}_{T-t}(x_t, y)}{\widetilde{q}_{T-t}(x_t, y) + (1 - \widetilde{q}_{T-t}(x_t, y)) \cdot \exp\left(-(e^{-2(T-t)} - e^{-(T-t)})(\|\mu_{y'}\|_2^2 - \|\mu_y\|_2^2)/2\right)} \\
\le& \frac{\widetilde{q}_{T-t}(x_t, y)}{\widetilde{q}_{T-t}(x_t, y) + (1 - \widetilde{q}_{T-t}(x_t, y)) \cdot \exp\left(-\Delta/8\right)},
\end{aligned}
\tag{22}
$$

where

$$
\widetilde{q}_{T-t}(x_t, y) = \frac{w_y}{w_y + \sum_{y'\neq y} w_{y'} \exp\left(e^{-(T-t)}\langle x_t, \mu_{y'} - \mu_y\rangle - e^{-(T-t)}(\|\mu_{y'}\|_2^2 - \|\mu_y\|_2^2)/2\right)}.
\tag{23}
$$

If $\exp(\langle x_t, \mu_y\rangle - \|\mu_y\|_2^2/2) = \max_{y'\in\mathcal{Y}} \exp(\langle x_t, \mu_{y'}\rangle - \|\mu_{y'}\|_2^2/2)$, then one can verify that

$$
\widetilde{q}_{T-t}(x_t, y) = \frac{w_y \exp\left(e^{-(T-t)}\langle x_t, \mu_y\rangle - e^{-2(T-t)}\|\mu_y\|_2^2/2\right)}{\sum_{y'\in\mathcal{Y}} w_{y'} \exp\left(e^{-(T-t)}\langle x_t, \mu_{y'}\rangle - e^{-2(T-t)}\|\mu_{y'}\|_2^2/2\right)} \le \mathcal{P}(x_t, y)
$$

Plugging this upper bound back into Eq. (22), we get

$$q_{T-t}(x_t, y) \leq \frac{\mathcal{P}(x_t, y)}{\mathcal{P}(x_t, y) + (1 - \mathcal{P}(x_t, y)) \cdot \exp(-\Delta/8)}. \tag{24}$$

On the other hand, if $\exp(\langle x_t, \mu_y \rangle - \|\mu_y\|_2^2/2) \neq \max_{y' \in \mathcal{Y}} \exp(\langle x_t, \mu_{y'} \rangle - \|\mu_{y'}\|_2^2/2)$, then one can verify that $\widetilde{q}_{T-t}(x_t, y) \leq w_y/(w_y + \min_{y' \neq y} w_{y'})$, which together with Eq. (22) further implies that

$$q_{T-t}(x_t, y) \leq \frac{w_y}{w_y + \min_{y' \neq y} w_{y'} \exp(-\Delta/8)}. \tag{25}$$

Putting together Eq. (24) and (25), we conclude that

$$q_{T-t}(x_t, y) \leq \max \left\{ G(\mathcal{P}(x_t, y)),\, G(w_y/(w_y + \min_{y' \neq y} w_{y'})) \right\}, \tag{26}$$

where $G(x) := x/(x + (1-x) \cdot \exp(-\Delta/8))$ is a function that maps $[0, 1]$ to $[0, 1]$. Taking the derivative of G, we see that for all $x \in [0, 1]$,

$$G'(x) = \frac{\exp(-\Delta/8)}{[x + (1-x) \cdot \exp(-\Delta/8)]^2} \in \big[\exp(-\Delta/8), \exp(\Delta/8) \big]. \tag{27}$$

Let $\xi_w := 1 - w_y/(w_y + \min_{y' \neq y} w_{y'}) > 0$. Note that $G(1) = 1$, hence by Eq. (27) we obtain that $1 - G(\mathcal{P}(x_t, y)) \geq \exp(-\Delta/8) \cdot (1 - \mathcal{P}(x_t, y))$ and $1 - G(1 - \xi_w) \geq \exp(-\Delta/8) \cdot \xi_w$. Substituting these bounds as well as the upper bound of Eq. (26) into Eq. (14), we are able to derive a lower bound for $\mathrm{d}\langle x_t, \mu_y - \mu_{y'} \rangle / \mathrm{d}t$:

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t} \langle x_t, \mu_y - \mu_{y'} \rangle \\ \geq & e^{-(T-t)} \Big(\|\mu_y\|_2^2 - \langle \mu_y, \mu_{y'} \rangle + \eta e^{-\Delta/8} (\|\mu_y - \mu_0\|_2^2 - 3\varepsilon) \cdot \min\{1 - \mathcal{P}(x_t, y),\, \xi_w\} \Big). \end{aligned} \tag{28}$$

Equivalently, we can write Eq. (28) as

$$\frac{\mathrm{d}}{\mathrm{d}t} \langle x_t - z_t, \mu_y - \mu_{y'} \rangle \geq e^{-(T-t)} \eta e^{-\Delta/8} (\|\mu_y - \mu_0\|_2^2 - 3\varepsilon) \cdot \min\{1 - \mathcal{P}(x_t, y),\, \xi_w\}.$$

Now suppose $\langle x_t, \mu_y - \mu_{y'} \rangle - \langle z_t, \mu_y - \mu_{y'} \rangle \in [0, \mathcal{U}]$ for all $t \in [0, T]$ and $y' \in \mathcal{Y}$. Using this bound, we get

$$\begin{aligned} \mathcal{P}(x_t, y) \leq & \frac{\mathcal{P}(z_t, y)}{\mathcal{P}(z_t, y) + (1 - \mathcal{P}(z_t, y)) \cdot \exp(-\mathcal{U})} \\ \leq & \frac{\max_{0 \leq t \leq T} \mathcal{P}(z_t, y)}{\max_{0 \leq t \leq T} \mathcal{P}(z_t, y) + (1 - \max_{0 \leq t \leq T} \mathcal{P}(z_t, y)) \cdot \exp(-\mathcal{U})} \\ = & 1 - \mathcal{F}\big(\max_{0 \leq t \leq T} \mathcal{P}(z_t, y), \mathcal{U} \big), \end{aligned} \tag{29}$$

where we let $\mathcal{F}(p, u) = (1-p)e^{-u}/(p + (1-p)e^{-u})$. Therefore, in order for such a $\mathcal{U} \in \mathbb{R}_{\geq 0}$ to serve as a valid upper bound, it is necessary to have

$$\mathcal{U} \geq \langle x_0 - z_0, \mu_y - \mu_{y'} \rangle + (1 - e^{-T}) \cdot \eta e^{-\Delta/8} (\|\mu_y - \mu_0\|_2^2 - 3\varepsilon) \cdot \min \Big\{ \mathcal{F}\big(\max_{0 \leq t \leq T} \mathcal{P}(z_t, y), \mathcal{U} \big),\, \xi_w \Big\}. \tag{30}$$

For any $\mathcal{U} \in \mathbb{R}_{\geq 0}$ that does not satisfy Eq. (30), we know that $\langle x_T - z_T, \mu_y - \mu_{y'} \rangle \geq \mathcal{U}$, hence

$$\mathcal{P}(x_T, y) \geq \frac{\mathcal{P}(z_T, y)}{\mathcal{P}(z_T, y) + (1 - \mathcal{P}(z_T, y)) \cdot \exp(-\mathcal{U})}.$$

This completes the proof of the first result of the theorem.

PROOF OF THE SECOND RESULT

We separately discuss two cases, depending on whether $\langle \mu_y, \mu_y - \mu_{y'} \rangle$ is non-negative for all $y' \in \mathcal{Y}$.

If $\langle \mu_y, \mu_y - \mu_{y'} \rangle \geq 0$ for all $y' \in \mathcal{Y}$, then by Eq. (14) we know that $t \mapsto \langle x_t, \mu_y - \mu_{y'} \rangle$ as a function of t is non-decreasing, which further implies that $\mathcal{P}(x_{t_1}, y) \geq \mathcal{P}(x_{t_2}, y)$ for all $T \geq t_1 \geq t_2 \geq 0$. We denote by $p(\eta)$ an upper bound for $\mathcal{P}(x_T, y)$ that implicitly depends on x_0. Following the analysis we have established to prove the first point (in particular, Eq. (31)), we have

$$q_{T-t}(x_t, y) \leq \max\left\{G(p(\eta)), G(1-\xi_w)\right\}.$$

Putting together the above upper bound and Eq. (14), we obtain that

$$\begin{aligned}&\langle x_T, \mu_y - \mu_{y'} \rangle - \langle x_0, \mu_y - \mu_{y'} \rangle \\ \geq &(1 - e^{-T}) \cdot \left(\langle \mu_y, \mu_y - \mu_{y'} \rangle + \eta \min\{1 - G(p(\eta)), 1 - G(1-\xi_w)\}(\|\mu_y - \mu_0\|_2^2 - 3\varepsilon) \right),\end{aligned} \tag{31}$$

where we recall that $\xi_w = 1 - w_y/(w_y + \min_{y' \neq y} w_{y'})$ and $G(x) = x/(x + (1-x)\cdot \exp(-\Delta/8))$. Note that $G(1) = 1$, hence by Eq. (27) we obtain that $1 - G(p(\eta)) \geq \exp(-\Delta/8) \cdot (1 - p(\eta))$ and $1 - G(1-\xi_w) \geq \exp(-\Delta/8)\xi_w$. Plugging this lower bound into Eq. (31), we get

$$\mathcal{P}(x_T, y) \geq \frac{\mathcal{P}(x_0, y)}{\mathcal{P}(x_0, y) + (1 - \mathcal{P}(x_0, y)) \cdot \exp\left(-C_0 - \eta C_1 \min\{1 - p(\eta), \xi_w\}\right)},$$

where $C_0 = \min_{y' \in \mathcal{Y}} (1 - e^{-T}) \langle \mu_y, \mu_y - \mu_{y'} \rangle$ and $C_1 = e^{-\Delta/8}(\|\mu_y - \mu_0\|_2^2 - 3\varepsilon)$. By definition we know $p(\eta) \geq \mathcal{P}(x_T, y)$, hence

$$\mathcal{P}(x_0, y)(1 - p(\eta)) \leq (1 - \mathcal{P}(x_0, y)) \cdot \exp\left(-C_0 - \eta C_1 \min\{1 - p(\eta), \xi_w\}\right). \tag{32}$$

When $\eta > 1$, we can write

$$1 - p(\eta) = \frac{-C_0 - \mathrm{logit}(\mathcal{P}(x_0, y)) + \delta_\eta \log \eta}{\eta C_1}, \tag{33}$$

where $\mathrm{logit}(p) = \log(p/(1-p))$, and $\delta_\eta > 0$. Plugging Eq. (33) into Eq. (32), we see that at least one of the following two inequalities hold:

$$\begin{aligned}\frac{-C_0 - \mathrm{logit}(\mathcal{P}(x_0, y)) + \delta_\eta \log \eta}{\eta C_1} &\leq \frac{1}{\eta^{\delta_\eta}}, \\ \frac{-C_0 - \mathrm{logit}(\mathcal{P}(x_0, y)) + \delta_\eta \log \eta}{\eta C_1} &\leq \frac{1 - \mathcal{P}(x_0, y)}{\mathcal{P}(x_0, y)} \cdot \exp(-C_0 - \eta C_1 \xi_w).\end{aligned}$$

Inspecting the above formulas, we see that for a sufficiently large η, it holds that $\delta_\eta < 1$, which implies that

$$p(\eta) \geq 1 - \frac{-C_0 - \mathrm{logit}(\mathcal{P}(x_0, y)) + \log \eta}{\eta C_1}. \tag{34}$$

On the other hand, if not all $\langle \mu_y, \mu_y - \mu_{y'} \rangle$ are non-negative, then we denote the smallest one by $-V_s = \langle \mu_y, \mu_y - \mu_{y_s} \rangle < 0$ for some $y_s \in \mathcal{Y}$. We shall choose η that is large enough such that $V_s < \eta e^{-\Delta/8} \xi_w (\|\mu_y - \mu_0\|_2^2 - 3\varepsilon)$. In this case, for all $y' \in \mathcal{Y}$, it holds that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} \langle x_t, \mu_y - \mu_{y'} \rangle \geq & e^{-T+t}(-V_s + \eta \min\{1 - G(\mathcal{P}(x_t, y)), 1 - G(1-\xi_w)\}(\|\mu_y - \mu_0\|_2^2 - 3\varepsilon)) \\ \geq & e^{-T+t}(-V_s + \eta e^{-\Delta/8} \min\{1 - \mathcal{P}(x_t, y), \xi_w\})(\|\mu_y - \mu_0\|_2^2 - 3\varepsilon).\end{aligned}$$

Therefore, if $1 - \mathcal{P}(x_t, y) \geq V_s e^{\Delta/8} \eta^{-1} (\|\mu_y - \mu_0\|_2^2 - 3\varepsilon)^{-1}$ for all $t \in [0, T]$, then $\langle x_t, \mu_y - \mu_{y'} \rangle$ as a function of t is non-decreasing on $[0, T]$, hence $\mathcal{P}(x_t, y)$ as a function of t is also non-decreasing on $[0, T]$. Following exactly the same route before, we are able to derive the lower bound as stated in Eq. (34). On the other hand, if $1 - \mathcal{P}(x_t, y) < V_s e^{\Delta/8} \eta^{-1} (\|\mu_y - \mu_0\|_2^2 - 3\varepsilon)^{-1}$ at some time point $t = t_*$, then for all $t \in [t_*, T]$, it is not hard to see that

$$1 - \mathcal{P}(x_t, y) \leq \frac{V_s e^{\Delta/8}}{\eta(\|\mu_y - \mu_0\|_2^2 - 3\varepsilon)}.$$

Putting together the above results, we conclude that for a sufficiently large η we always have Eq. (34). The proof is complete.

A.3 Proof of Theorem 3.7

We initiate the proof by presenting an SDE comparison theorem. Lemma A.1 is adapted from Theorem 3.1 of (Zhu, 2010).

Lemma A.1 (SDE comparison theorem)**.** *Consider the following two m-dimensional SDEs defined on $[0, T]$:*

$$\begin{aligned} X_t^1 &= x^1 + \int_0^t b_1(s, X_s^1)\mathrm{d}s + \int_0^t \sigma_1(s, X_s^1)\mathrm{d}W_s, \\ X_t^2 &= x^2 + \int_0^t b_2(s, X_s^2)\mathrm{d}s + \int_0^t \sigma_2(s, X_s^2)\mathrm{d}W_s. \end{aligned}$$

We assume the following conditions:

1. *$b(t, x), \sigma(t, x)$ are continuous in (t, x),*
2. *There exists a sufficiently large constant $\mu > 0$, such that for all $x, x' \in \mathbb{R}^m$ and $t \in [0, T]$, it holds that*

$$\begin{aligned} &\|b(t, x) - b(t, x')\|_2 + \|\sigma(t, x) - \sigma(t, x')\|_2 \le \mu\|x - x'\|_2, \\ &\|b(t, x)\|_2 + \|\sigma(t, x)\|_2 \le \mu(1 + \|x\|_2). \end{aligned}$$

Then the following are equivalent:

(i) For any $t \in [0, T]$ and $x^1, x^2 \in \mathbb{R}^m$ such that $x^1 \ge x^2$, almost surely we have $X_t^1 \ge X_t^2$ for all $t \in [0, T]$.

(ii) $\sigma^1 \equiv \sigma^2$, and for any $t \in [0, T]$, $k = 1, 2, \cdots, m$,

$$\begin{cases} (a) & \sigma_k^1 \text{ depends only on } x_k, \\ (b) & \text{for all } x', \delta^k x \in \mathbb{R}^m, \text{ such that } \delta^k x \ge 0, (\delta^k x)_k = 0, \\ & \quad b_k^1(t, \delta^k x + x') \ge b_k^2(t, x'). \end{cases}$$

We then prove the theorem. To this end, we establish the subsequent lemma. Note that Theorem 3.7 follows straightforwardly from Lemma A.2.

Lemma A.2. *We assume the conditions of Theorem 3.7. Then for all $y' \in \mathcal{Y}$, almost surely we have $\langle x_t, \mu_y - \mu_{y'}\rangle \ge \langle z_t, \mu_y - \mu_{y'}\rangle$ for all $t \in [0, T]$.*

Proof of Lemma A.2. Note that

$$\begin{aligned} &\mathrm{d}\langle \bar{x}_t, \mu_y - \mu_{y'}\rangle \\ &= \Bigg[-\langle \bar{x}_t, \mu_y - \mu_{y'}\rangle + 2e^{-(T-t)}(1 + \eta - \eta q_{T-t}(\bar{x}_t, y))\|\mu_y\|_2^2 - 2\eta e^{-(T-t)} \sum_{y'' \ne y} q_{T-t}(\bar{x}_t, y'')\langle \mu_y, \mu_{y''}\rangle \\ &\quad -2e^{-(T-t)}(1 + \eta - \eta q_{T-t}(\bar{x}_t, y))\langle \mu_y, \mu_{y'}\rangle + 2\eta e^{-(T-t)} \sum_{y'' \ne y} q_{T-t}(\bar{x}_t, y'')\langle \mu_{y'}, \mu_{y''}\rangle\Bigg] \mathrm{d}t \\ &\quad + \sqrt{2}\langle \mathrm{d}B_t, \mu_y - \mu_{y'}\rangle \\ &= \Big[-\langle \bar{x}_t, \mu_y - \mu_{y'}\rangle + 2e^{-(T-t)}\|\mu_y\|_2^2 - 2e^{-(T-t)}\langle \mu_y, \mu_{y'}\rangle + 2\eta e^{-(T-t)}(1 - q_{T-t}(\bar{x}_t, y))\|\mu_y - \mu_0\|_2^2 \\ &\quad +2\eta e^{-(T-t)} q_{T-t}(\bar{x}_t, y')\|\mu_{y'} - \mu_0\|_2^2 + \bar{\mathcal{E}}_t\Big] + \sqrt{2}\langle \mathrm{d}B_t, \mu_y - \mu_{y'}\rangle \\ &\ge \Big[-\langle \bar{x}_t, \mu_y - \mu_{y'}\rangle + 2e^{-(T-t)}\|\mu_y\|_2^2 - 2e^{-(T-t)}\langle \mu_y, \mu_{y'}\rangle + 2\eta e^{-(T-t)}(1 - q_{T-t}(\bar{x}_t, y))(\|\mu_y - \mu_0\|_2^2 - 3\varepsilon)\Big] \mathrm{d}t \\ &\quad + \sqrt{2}\langle \mathrm{d}B_t, \mu_y - \mu_{y'}\rangle, \end{aligned} \tag{35}$$

where $\bar{\mathcal{E}}_t$ is a function of $(\bar{x}_t, t)$, and $|\bar{\mathcal{E}}_t| \le 6\eta e^{-(T-t)}(1 - q_{T-t}(\bar{x}_t, y))\varepsilon$. Note that the unguided process $(\bar{z}_t)_{0 \le t \le T}$ satisfies the following SDE

$$\mathrm{d}\langle \bar{z}_t, \mu_y - \mu_{y'}\rangle = \Big[-\langle \bar{z}_t, \mu_y - \mu_{y'}\rangle + 2e^{-(T-t)}\|\mu_y\|_2^2 - 2e^{-(T-t)}\langle \mu_y, \mu_{y'}\rangle\Big] \mathrm{d}t + \sqrt{2}\langle \mathrm{d}B_t, \mu_y - \mu_{y'}\rangle. \tag{36}$$

Lemma A.2 then follows as a straightforward consequence of Lemma A.1.

□

A.4 Proof of Theorem 3.8

Plugging Eq. (26) and (27) into the last line of Eq. (35), we obtain

$$
\begin{aligned}
&\mathrm{d}\langle \bar{x}_t, \mu_y - \mu_{y'} \rangle \\
\geq &\Big[-\langle \bar{x}_t, \mu_y - \mu_{y'} \rangle + 2e^{-(T-t)} \Big(\|\mu_y\|_2^2 - \langle \mu_y, \mu_{y'} \rangle + \eta e^{-\Delta/8} \min\{1 - \mathcal{P}(x_t, y), \xi_w\} (\|\mu_y - \mu_0\|_2^2 - 3\varepsilon) \Big) \Big] \,\mathrm{d}t \\
&+ \sqrt{2} \langle \mathrm{d}B_t, \mu_y - \mu_{y'} \rangle. \qquad (37)
\end{aligned}
$$

Invoking the method of integrating factors, we see that

$$
\begin{aligned}
&\mathrm{d}\left[e^t \langle \bar{x}_t, \mu_y - \mu_{y'} \rangle \right] \\
\geq &2e^{-T+2t} \Big(\|\mu_y\|_2^2 - \langle \mu_y, \mu_{y'} \rangle + \eta e^{-\Delta/8} \min\{1 - \mathcal{P}(x_t, y), \xi_w\} (\|\mu_y - \mu_0\|_2^2 - 3\varepsilon) \Big) \mathrm{d}t + \sqrt{2} e^t \langle \mathrm{d}B_t, \mu_y - \mu_{y'} \rangle.
\end{aligned}
$$

Note that

$$
\mathrm{d}\left[e^t \langle \bar{z}_t, \mu_y - \mu_{y'} \rangle \right] = 2e^{-T+2t} \Big(\|\mu_y\|_2^2 - \langle \mu_y, \mu_{y'} \rangle \Big) \mathrm{d}t + \sqrt{2} e^t \langle \mathrm{d}B_t, \mu_y - \mu_{y'} \rangle.
$$

Combining the above two equations, we obtain that

$$
\mathrm{d}\left[e^t \langle \bar{x}_t - \bar{z}_t, \mu_y - \mu_{y'} \rangle \right] \geq 2\eta e^{-T+2t-\Delta/8} \min\{1 - \mathcal{P}(x_t, y), \xi_w\} (\|\mu_y - \mu_0\|_2^2 - 3\varepsilon) \mathrm{d}t. \qquad (38)
$$

By assumption $\langle \bar{x}_0 - \bar{z}_0, \mu_y - \mu_{y'} \rangle \geq 0$, hence $\langle \bar{x}_t - \bar{z}_t, \mu_y - \mu_{y'} \rangle \geq 0$ for all $t \in [0, T]$. Suppose we have $e^t \langle \bar{x}_t, \mu_y - \mu_{y'} \rangle - e^t \langle \bar{z}_t, \mu_y - \mu_{y'} \rangle \in [0, \mathcal{U}]$ for all $t \in [0, T]$ and $y' \in \mathcal{Y}$. Then from Eq. (29) we know that for all $t \in [0, T]$,

$$
\mathcal{P}(\bar{x}_t, y) \leq 1 - \mathcal{F}\big(\max_{0 \leq t \leq T} \mathcal{P}(\bar{z}_t, y), \mathcal{U} \big). \qquad (39)
$$

Using Eq. (38) and (39), we see that in order for $\mathcal{U}$ to be a valid upper bound, we must have

$$
\mathcal{U} \geq \langle \bar{x}_0 - \bar{z}_0, \mu_y - \mu_{y'} \rangle + \eta (e^T - e^{-T}) e^{-\Delta/8} \min\big\{ \mathcal{F}\big(\max_{0 \leq t \leq T} \mathcal{P}(\bar{z}_t, y), \mathcal{U} \big), \xi_w \big\} (\|\mu_y - \mu_0\|_2^2 - 3\varepsilon). \qquad (40)
$$

For any $\mathcal{U}$ that does not satisfy Eq. (40), we know that there exists $t \in [0, T]$, such that $e^t \langle \bar{x}_t - \bar{z}_t, \mu_y - \mu_{y'} \rangle \geq \mathcal{U}$, hence $\langle \bar{x}_T - \bar{z}_T, \mu_y - \mu_{y'} \rangle \geq e^{-T} \mathcal{U}$. As a consequence, we have

$$
\mathcal{P}(\bar{x}_T, y) \geq \frac{\mathcal{P}(\bar{z}_T, y)}{\mathcal{P}(\bar{z}_T, y) + (1 - \mathcal{P}(\bar{z}_T, y)) \cdot \exp(-e^{-T} \mathcal{U})}. \qquad (41)
$$

Setting $\bar{\mathcal{U}} = e^{-T} \mathcal{U}$ completes the proof of the first result.

As for the proof of the convergence rate, note that if we set $e^{-\mathcal{U}} = \eta^{-1} (\log \eta)^2$, then as $\eta \to \infty$, the left hand side of Eq. (40) is of order $O(\log \eta)$, while the right hand side of Eq. (40) is of order $O(\eta \wedge (\log \eta)^2)$. Hence, for a large enough η Eq. (40) is not satisfied. Plugging such $\mathcal{U}$ into Eq. (41), we conclude that $\mathcal{P}(\bar{x}_T, y) \geq 1 - O(\eta^{-e^{-T}} (\log \eta)^{2e^{-T}})$.

A.5 Proof of Theorem 3.9

Proof of the first claim

Inspecting Eq. (21), (15) and setting $\mu_0 = (\mu_1 + \mu_2)/2$, $\mu = \mu_1 - \mu_0$ therein, we have

$$
2 \frac{\mathrm{d}}{\mathrm{d}t} \langle x_t, \mu \rangle = e^{-(T-t)} \|\mu_1\|_2^2 - e^{-(T-t)} \langle \mu_1, \mu_2 \rangle + 4\eta e^{-(T-t)} (1 - q_{T-t}(x_t, 1)) \|\mu\|_2^2, \qquad (42)
$$

$$
2 \frac{\mathrm{d}}{\mathrm{d}t} \langle z_t, \mu \rangle = e^{-(T-t)} \|\mu_1\|_2^2 - e^{-(T-t)} \langle \mu_1, \mu_2 \rangle. \qquad (43)
$$

Applying the ODE comparison theorem (Lemma 3.4), we conclude that $\langle x_t, \mu\rangle \geq \langle z_t, \mu\rangle$ for all $t \in [0, T]$. The first claim of the lemma then immediately follows, as in this case

$$\mathcal{P}(x_t, 1) = \frac{w_1 \exp(\langle x_t, \mu\rangle - \|\mu_1\|_2^2/2)}{w_1 \exp(\langle x_t, \mu\rangle - \|\mu_1\|_2^2/2) + w_2 \exp(-\langle x_t, \mu\rangle - \|\mu_2\|_2^2/2)},$$
$$\mathcal{P}(z_t, 1) = \frac{w_1 \exp(\langle z_t, \mu\rangle - \|\mu_1\|_2^2/2)}{w_1 \exp(\langle z_t, \mu\rangle - \|\mu_1\|_2^2/2) + w_2 \exp(-\langle z_t, \mu\rangle - \|\mu_2\|_2^2/2)}.$$

PROOF OF THE SECOND CLAIM

Similar to the derivation of Eq. (22), we conclude that

$$q_{T-t}(x_t, 1) \leq \frac{\widetilde{q}_{T-t}(x_t, 1)}{\widetilde{q}_{T-t}(x_t, 1) + (1 - \widetilde{q}_{T-t}(x_t, 1)) \cdot \exp(-\Delta_1/8)}, \tag{44}$$

where we recall that $\widetilde{q}_{T-t}$ is defined in Eq. (23), and $\Delta_1 = |\|\mu_1\|_2^2 - \|\mu_2\|_2^2|$. If $\exp(\langle x_t, \mu_1\rangle - \|\mu_1\|_2^2/2) \geq \exp(\langle x_t, \mu_2\rangle - \|\mu_2\|_2^2/2)$, then

$$\widetilde{q}_{T-t}(x_t, 1) \leq \mathcal{P}(x_t, 1).$$

Plugging the above inequality into Eq. (44), we obtain that

$$q_{T-t}(x_t, 1) \leq \frac{\mathcal{P}(x_t, 1)}{\mathcal{P}(x_t, 1) + (1 - \mathcal{P}(x_t, 1)) \cdot \exp(-\Delta_1/8)}. \tag{45}$$

On the other hand, if $\exp(\langle x_t, \mu_1\rangle - \|\mu_1\|_2^2/2) \leq \exp(\langle x_t, \mu_2\rangle - \|\mu_2\|_2^2/2)$, then $\widetilde{q}_{T-t}(x_t, 1) \leq w_1$. Putting together this upper bound, Eq. (44) and (45), we conclude that

$$q_{T-t}(x_t, 1) \leq \max\{G_1(\mathcal{P}(x_t, 1)), G_1(w_1)\},$$

where $G_1(x) := x/(x + (1 - x) \cdot \exp(-\Delta_1/8))$ maps $[0, 1]$ to $[0, 1]$. Taking the derivative of G_1, we see that for all $x \in [0, 1]$,

$$G_1'(x) = \frac{\exp(-\Delta_1/8)}{[x + (1 - x) \cdot \exp(-\Delta_1/8)]^2} \in \big[\exp(-\Delta_1/8), \exp(\Delta_1/8)\big]. \tag{46}$$

Observe that $G_1(1) = 1$, then by Eq. (46) we have

$$1 - G_1(\mathcal{P}(x_t, 1)) \geq e^{-\Delta_1/8}(1 - \mathcal{P}(x_t, 1)), \qquad 1 - G_1(w_1) \geq e^{-\Delta_1/8}(1 - w_1),$$

which further implies that

$$1 - q_{T-t}(x_t, 1) \geq e^{-\Delta_1/8} \min\{1 - \mathcal{P}(x_t, 1), 1 - w_1\}. \tag{47}$$

Plugging the lower bound in Eq. (47) into Eq. (42) and (43), we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle x_t - z_t, \mu\rangle \geq 2\eta e^{-T+t-\Delta_1/8} \min\{1 - \mathcal{P}(x_t, 1), 1 - w_1\}\|\mu\|_2^2. \tag{48}$$

The above equation together with the assumption $\langle x_0 - z_0, \mu\rangle \geq 0$ implies that $\langle x_t - z_t, \mu\rangle \geq 0$ for all $t \in [0, T]$. Now suppose $2\langle x_t - z_t, \mu\rangle \in [0, \mathcal{U}]$ for all $t \in [0, T]$. Similar to the derivation of Eq. (29), we conclude that

$$\mathcal{P}(x_t, 1) \leq 1 - \mathcal{F}\big(\max_{0 \leq t \leq T} \mathcal{P}(z_t, 1), \mathcal{U}\big). \tag{49}$$

Plugging Eq. (49) into Eq. (48), we see that in order for $\mathcal{U}$ to be a valid upper bound, we must have

$$\mathcal{U} \geq 2\langle x_0 - z_0, \mu\rangle + 4\eta e^{-\Delta_1/8}\|\mu\|_2^2(1 - e^{-T}) \min\big\{\mathcal{F}\big(\max_{0 \leq t \leq T} \mathcal{P}(z_t, 1), \mathcal{U}\big), 1 - w_1\big\}. \tag{50}$$

If $\mathcal{U}$ does not satisfy Eq. (50), then $2\langle x_T - z_T, \mu\rangle \geq \mathcal{U}$, hence

$$\mathcal{P}(x_T, 1) \geq \frac{\mathcal{P}(z_T, 1)}{\mathcal{P}(z_T, 1) + (1 - \mathcal{P}(z_T, 1)) \cdot \exp(-\mathcal{U})}. \tag{51}$$

The proof of the first result is complete. We then prove the result regarding the convergence rate as $\eta \to \infty$. To this end, we set $e^{-\mathcal{U}} = \eta^{-1}(\log \eta)^2$. For such $\mathcal{U}$, the left hand side of Eq. (50) is of order $O(\log \eta)$, while the right hand side of Eq. (50) is of order $O(\eta \wedge (\log \eta)^2)$. Therefore, for a sufficiently large η, Eq. (50) does not hold. Plugging such $\mathcal{U}$ into Eq. (51), we deduce that $\mathcal{P}(x_T, 1) \geq 1 - O(\eta^{-1}(\log \eta)^2)$ as $\eta \to \infty$.

A.6 Proof of Theorem 3.10

PROOF OF THE FIRST CLAIM

Similar to the proof of Theorem 3.9, we set $\mu_0 = (\mu_1 + \mu_2)/2$ and $\mu = \mu_1 - \mu_0$. Following the derivation of Eq. (35) and (36), we obtain

$$\begin{aligned} 2\mathrm{d}\langle \bar{x}_t, \mu \rangle &= \Big[-2\langle \bar{x}_t, \mu \rangle + 2e^{-(T-t)} \|\mu_1\|_2^2 - 2e^{-(T-t)} \langle \mu_1, \mu_2 \rangle + 8\eta e^{-(T-t)} (1 - q_{T-t}(\bar{x}_t, 1)) \|\mu\|_2^2 \Big] \, \mathrm{d}t \\ &\quad + 2\sqrt{2} \langle \mathrm{d}B_t, \mu \rangle, \\ 2\mathrm{d}\langle \bar{z}_t, \mu \rangle &= \Big[-2\langle \bar{z}_t, \mu \rangle + 2e^{-(T-t)} \|\mu_1\|_2^2 - 2e^{-(T-t)} \langle \mu_1, \mu_2 \rangle \Big] \, \mathrm{d}t + 2\sqrt{2} \langle \mathrm{d}B_t, \mu \rangle. \end{aligned} \tag{52}$$

Note that $q_{T-t}(\bar{x}_t, 1)$ depends on $\bar{x}_t$ only through $\langle \bar{x}_t, \mu \rangle$, hence both equations listed above represent an SDE. Then, we may leverage the SDE comparison theorem (Lemma A.1) to deduce that almost surely, $\langle \bar{x}_t, \mu \rangle \geq \langle \bar{z}_t, \mu \rangle$ for all $t \in [0, T]$. This completes the proof of the first claim.

PROOF OF THE SECOND CLAIM

Plugging Eq. (47) into Eq. (52), we see that

$$2\mathrm{d}\langle \bar{x}_t - \bar{z}_t, \mu \rangle \geq \Big[-2\langle \bar{x}_t - \bar{z}_t, \mu \rangle + 8\eta \|\mu\|_2^2 e^{-T+t-\Delta_1/8} \min\{1 - \mathcal{P}(\bar{x}_t, 1), w_2\} \Big] \, \mathrm{d}t.$$

Multiplying both sides above by e^t, we get

$$\mathrm{d} \left[2e^t \langle \bar{x}_t - \bar{z}_t, \mu \rangle \right] \geq 8\eta \|\mu\|_2^2 e^{-T+2t-\Delta_1/8} \min\{1 - \mathcal{P}(\bar{x}_t, 1), w_2\} \mathrm{d}t. \tag{53}$$

Since by assumption $\langle \bar{x}_0 - \bar{z}_0, \mu \rangle \geq 0$, we then conclude that almost surely we have $\langle \bar{x}_t - \bar{z}_t, \mu \rangle \geq 0$ for all $t \in [0, T]$. If we assume $2e^t \langle \bar{x}_t - \bar{z}_t, \mu \rangle \in [0, \mathcal{U}]$ for all $t \in [0, T]$, then it holds that $2\langle \bar{x}_t - \bar{z}_t, \mu \rangle \leq \mathcal{U}$ for all $t \in [0, T]$. Following the derivation of Eq. (39), we have

$$1 - \mathcal{P}(\bar{x}_t, 1) \geq \mathcal{F}\big(\max_{0 \leq t \leq T} \mathcal{P}(\bar{z}_t, 1), \mathcal{U} \big). \tag{54}$$

Putting together Eq. (53) and (54), we see that for $\mathcal{U}$ to serve as a valid upper bound, we must have

$$\mathcal{U} \geq 2\langle \bar{x}_0 - \bar{z}_0, \mu \rangle + 4\eta \|\mu\|_2^2 e^{-\Delta_1/8} (e^T - e^{-T}) \min\Big\{ \mathcal{F}\big(\max_{0 \leq t \leq T} \mathcal{P}(\bar{z}_t, 1), \mathcal{U} \big), w_2 \Big\}. \tag{55}$$

If Eq. (55) is not satisfied, then for such $\mathcal{U}$ we have $2\langle \bar{x}_T - \bar{z}_T, \mu \rangle \geq e^{-T} \mathcal{U}$, and

$$\mathcal{P}(\bar{x}_T, 1) \geq \frac{\mathcal{P}(\bar{z}_T, 1)}{\mathcal{P}(\bar{z}_T, 1) + (1 - \mathcal{P}(\bar{z}_T, 1)) \cdot \exp(-e^{-T} \mathcal{U})}. \tag{56}$$

Setting $\bar{\mathcal{U}} = e^{-T} \mathcal{U}$ completes the proof of the first bound.

To prove the convergence rate, we simply set $e^{-\mathcal{U}} = \eta^{-1} (\log \eta)^2$. As $\eta \to \infty$, the left hand side of Eq. (55) is of order $O(\log \eta)$, while the right hand side of Eq. (55) is of order $O(\eta \wedge (\log \eta)^2)$. For a large enough η, we see that Eq. (55) does not hold. Plugging such $\mathcal{U}$ into Eq. (56), we conclude that $\mathcal{P}(\bar{x}_T, 1) = 1 - O(\eta^{-e^{-T}} (\log \eta)^{2e^{-T}})$.

B Proofs related to diversity reduction

This section contains proofs related to diversity reduction. We present in Appendix B.1 a heuristic derivation of Theorem 4.2 based on the Fokker-Planck equation, and leave the establishment of a rigorous procedure to the remaining sections. In Appendix B.2, we demonstrate the existence of probability density functions $Q(\cdot, t)$ and $Q_0(\cdot, t)$ for all $t \in [0, T]$.

B.1 Derivation of Theorem 4.2 via the Fokker-Planck equation

We provide in this section a non-rigorous derivation of Theorem 4.2 via the Fokker-Planck equation. This part serves as a motivation of our theorem, and a rigorous proof can be found in Appendix B.3 instead.

Leveraging the Fokker–Planck equation (Lemma 4.1), on $[0, T]$ we have

$$\begin{aligned}
\frac{\partial}{\partial t} Q(t, x) &= -\sum_{i=1}^{d} \frac{\partial}{\partial x_i}\left[Q(t, x) \cdot \big(x_i + \frac{\partial}{\partial x_i} \log p_{T-t}(x, y) + \eta \frac{\partial}{\partial x_i} \log p_{T-t}(y \mid x)\big)\right], \\
\frac{\partial}{\partial t} Q_0(t, x) &= -\sum_{i=1}^{d} \frac{\partial}{\partial x_i}\left[Q_0(t, x) \cdot \big(x_i + \frac{\partial}{\partial x_i} \log p_{T-t}(x, y)\big)\right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{\partial}{\partial t} H(t) &= -\int \frac{\partial}{\partial t} Q(t, x) \log Q(t, x) \mathrm{d}x - \int \frac{\partial}{\partial t} Q(t, x) \mathrm{d}x \\
&\overset{(i)}{=} \sum_{i=1}^{d} \int \frac{\partial}{\partial x_i}\left[Q(t, x) \cdot \big(x_i + \frac{\partial}{\partial x_i} \log p_{T-t}(x, y) + \eta \frac{\partial}{\partial x_i} \log p_{T-t}(y \mid x)\big)\right] \log Q(t, x) \mathrm{d}x \\
&= \sum_{i=1}^{d} \int \log Q(t, x) \frac{\partial}{\partial x_i} Q(t, x) \cdot \left[x_i + \frac{\partial}{\partial x_i} \log p_{T-t}(x, y) + \eta \frac{\partial}{\partial x_i} \log p_{T-t}(y \mid x)\right] \mathrm{d}x \\
&\quad + \sum_{i=1}^{d} \int Q(t, x) \log Q(t, x) \cdot \left[1 + \frac{\partial^2}{\partial^2 x_i} \log p_{T-t}(x, y) + \eta \frac{\partial^2}{\partial^2 x_i} \log p_{T-t}(y \mid x)\right] \mathrm{d}x \\
&\overset{(ii)}{=} \sum_{i=1}^{d} \int \left(1 + \frac{\partial^2}{\partial^2 x_i} \log p_{T-t}(x, y) + \eta \frac{\partial^2}{\partial^2 x_i} \log p_{T-t}(y \mid x)\right) Q(t, x) \mathrm{d}x,
\end{aligned}$$

where (i) is because $\int Q(t, x) \mathrm{d}x \equiv 1$, and (ii) is via integration by parts. Applying a similar procedure to the diffusion model without guidance, we obtain

$$\frac{\partial}{\partial t} H_0(t) = \sum_{i=1}^{d} \int \left(1 + \frac{\partial^2}{\partial^2 x_i} \log p_{T-t}(x, y)\right) Q_0(t, x) \mathrm{d}x.$$

Note that

$$\begin{aligned}
&\sum_{i=1}^{d} \frac{\partial^2}{\partial^2 x_i} \log p_{T-t}(x, y) = -\operatorname{tr}\left[\Sigma_{T-t}^{-1}\right], \\
&\sum_{i=1}^{d} \frac{\partial^2}{\partial^2 x_i} \log p_{T-t}(y \mid x) = -\operatorname{tr}\left[\sum_{y' \in \mathcal{Y}} q_{T-t}(y' \mid x) \Sigma_{T-t}^{-1} \mu_{y'} \mu_{y'}^{\top} \Sigma_{T-t}^{-1} - v v^{\top}\right], \\
&v = \sum_{y' \in \mathcal{Y}} q_{T-t}(y' \mid x) \Sigma_{T-t}^{-1} \mu_{y'}.
\end{aligned}$$

As a consequence, we have $\sum_{i=1}^{d} \frac{\partial^2}{\partial^2 x_i} \log p_{T-t}(y \mid x) \leq 0$. Putting together this result and the ODE comparison theorem (Lemma 3.4), we obtain the desired result. However, we emphasize that the above derivation is non-rigorous. For example, it is unclear whether the Fokker-Planck equation has a solution, and also the exchange of integration and differentiation is unjustified.

B.2 Existence of probability density functions

In this section, we justify the existence of probability density functions. Namely, we establish the following lemma.

Lemma B.1. *We assume the conditions of Theorem 4.2. Then $Q(t, \cdot)$ and $Q_0(t, \cdot)$ exist for all $t \in [0, T]$.*

We prove Lemma B.1 in the remainder of this section. We separately discuss the guided process and the unguided process below.

PROOF FOR Q_0

We first show that z_t has a probability density function for all $t \in [0,T]$. Observe that $(z_t)_{0\leq t\leq T}$ is a solution to the following ODE:

$$\frac{\mathrm{d}z_t}{\mathrm{d}t} = A(t)z_t + b(t), \tag{57}$$

where the symmetric matrix $A(t) \in \mathbb{R}^{d\times d}$ and the vector $b(t) \in \mathbb{R}^d$ depend only on t. In addition, $A(t)A(s) = A(s)A(t)$ for all $t, s \in [0,T]$. Solving Eq. (57), we conclude that

$$z_t = \exp\left(\int_0^t A(s)\mathrm{d}s\right) z_0 + \int_0^t \exp\left(\int_s^t A(i)\mathrm{d}i\right) b(s)\mathrm{d}s.$$

The matrix $\exp\left(\int_0^t A(s)\mathrm{d}s\right)$ is non-degenerate. By assumption, z_0 has a probability density function with respect to the Lebesgue measure. Therefore, z_t also has a density. The proof is complete.

PROOF FOR Q

We then prove the lemma for the guided process x_t. Inspecting Eq. (10) and applying the triangle inequality, we obtain

$$\begin{aligned}\frac{\mathrm{d}\|x_t\|_2}{\mathrm{d}t} &\geq -\|I_d - \Sigma_{T-t}^{-1}\|_{\mathrm{op}}\|x_t\|_2 - \|\Sigma_{T-t}^{-1}\|_{\mathrm{op}}\|\mu_y\|_2 - 2\eta\|\Sigma_{T-t}^{-1}\|_{\mathrm{op}} \sup_{y'\in\mathcal{Y}} \|\mu_{y'}\|_2 \\ &\geq -\Big(1 + [\sigma_{\min}(\Sigma)\wedge 1]^{-1}\Big)\|x_t\|_2 - [\sigma_{\min}(\Sigma)\wedge 1]^{-1}\|\mu_y\|_2 - 2\eta[\sigma_{\min}(\Sigma)\wedge 1]^{-1} \sup_{y'\in\mathcal{Y}} \|\mu_{y'}\|_2,\end{aligned}$$

which by Lemma 3.4 further implies that

$$\|x_t\|_2 \geq e^{-Ct}\|x_0\|_2 - \frac{D}{C}\cdot(1-e^{-Ct}), \tag{58}$$

where $C = (1+[\sigma_{\min}(\Sigma)\wedge 1]^{-1})$ and $D = [\sigma_{\min}(\Sigma)\wedge 1]^{-1}\|\mu_y\|_2 + 2\eta[\sigma_{\min}(\Sigma)\wedge 1]^{-1}\sup_{y'\in\mathcal{Y}}\|\mu_{y'}\|_2$. Now we consider the set of initial values that lead to x_t:

$$\mathcal{I}_t(x_t) := \{x_0 \in \mathbb{R}^d : \text{ ODE (10) with initial value } x_0 \text{ has value } x_t \text{ at time } t\}.$$

Examining Eq. (58), we conclude that $\mathcal{I}_t(x_t) \subseteq \mathcal{B}(f_t(\|x_t\|_2))$, where $\mathcal{B}(r)$ stands for the ball in $\mathbb{R}^d$ that has radius r and is centered at the origin, and $f_t(r) = e^{Ct}r + C^{-1}D(e^{Ct}-1)$.

For the sake of simplicity, we rewrite Eq. (10) as $\mathrm{d}x_t = F(x_t,t)\mathrm{d}t$. For any $\delta > 0$, we consider the approximation to the ODE defined in Eq. (10) that has step size δ:

$$\widehat{x}_{k\delta}^{\delta} = \widehat{x}_{(k-1)\delta}^{\delta} + \delta F(\widehat{x}_{(k-1)\delta}^{\delta}, (k-1)\delta), \qquad \widehat{x}_0^{\delta} = x_0.$$

For $t \in [(k-1)\delta, k\delta]$, we compute $\widehat{x}_t^{\delta}$ by linearly interpolating $\widehat{x}_{(k-1)\delta}^{\delta}$ and $\widehat{x}_{k\delta}^{\delta}$. To simplify analysis, we consider only δ that takes the form T/K for $K \in \mathbb{N}_+$. Taking the Jacobian matrix of $F(x,t)$ with respect to x, we get

$$\begin{aligned}&\nabla_x F(x,t) = I_d - \Sigma_{T-t}^{-1} - \eta e^{-T+t}\Omega_{T-t}(x), \\ &\Omega_{T-t}(x) = \sum_{y'\in\mathcal{Y}} q_{T-t}(x,y')\Sigma_{T-t}^{-1}\mu_{y'}\mu_{y'}^{\top}\Sigma_{T-t}^{-1} - \Big(\sum_{y'\in\mathcal{Y}} q_{T-t}(x,y')\Sigma_{T-t}^{-1}\mu_{y'}\Big)\Big(\sum_{y'\in\mathcal{Y}} q_{T-t}(x,y')\Sigma_{T-t}^{-1}\mu_{y'}\Big)^{\top}.\end{aligned}$$

Therefore, we conclude that $F(\cdot,t)$ is Lipschitz continuous in its first argument, and the Lipschitz constant is uniformly bounded for all $t \in [0,T]$. In addition, F is continuous in its second argument. Leveraging a standard Gronwall type argument, we obtain that for all $r > 0$,

$$\limsup_{\delta\to 0^+} \sup_{x_0\in\mathcal{B}(r)} \Big\{\sup_{0\leq t\leq T} \|\widehat{x}_t^{\delta} - x_t\|_2\Big\} = 0. \tag{59}$$

We can compute the Jacobian of the mapping $x_0 \mapsto \widehat{x}_t^\delta$. To simplify presentation, here we let $t = k\delta$. We comment that the treatment for general $t \in [0, T]$ is similar, and we leave the homework to interested readers. We denote the Jacobian of this mapping $x_0 \mapsto \widehat{x}_t^\delta$ by $J_{0\to k\delta}^\delta(x_0) \in \mathbb{R}^{d\times d}$. Observe that

$$J_{0\to k\delta}^\delta(x_0) = \prod_{i=0}^{k-1} \Big(I_d + \delta(I_d - \Sigma_{T-i\delta}^{-1} - \eta e^{-T+i\delta}\Omega_{T-i\delta}(\widehat{x}_{i\delta}^\delta)) \Big),$$

where $\prod_{i=0}^{k-1} A_i := A_{k-1}A_{k-2}\cdots A_0$. For a sufficiently small δ we see that $J_{0\to k\delta}^\delta(x_0)$ is non-degenerate for all k and x_0. In addition, one can verify that for fixed $T, r > 0$ ($T = K\delta$), it holds that

$$\begin{aligned}
&\lim_{\delta\to 0^+} \sup_{x_0\in\mathcal{B}(r), k\in\{0\}\cup[K]} \big\|J_{0\to k\delta}^\delta(x_0) - J_{0\to k\delta}(x_0)\big\|_{\mathrm{op}} = 0, \\
&J_{0\to k\delta}(x_0) = \exp\Big(\int_0^{k\delta} (I_d - \Sigma_{T-t}^{-1} - \eta e^{-T+t}\Omega_{T-t}(x_t))\mathrm{d}t \Big) \in \mathbb{R}^{d\times d}.
\end{aligned} \tag{60}$$

We write $x_t = G_t(x_0)$ and $\widehat{x}_t^\delta = \widehat{G}_t^\delta(x_0)$. By Eq. (59) we have $\lim_{\delta\to 0^+}\sup_{x_0\in\mathcal{B}(r), 0\le t\le T} \|G_t(x_0) - \widehat{G}_t^\delta(x_0)\|_2 = 0$. Next, we prove that $\nabla_{x_0} G_t(x_0) = J_{0\to t}(x_0)$. To this end, it suffice to show

$$\big\|G_t(x_0 + x') - G_t(x_0) - J_{0\to t}(x_0)x'\big\|_2 = o(\|x'\|_2).$$

By triangle inequality,

$$\begin{aligned}
&\big\|G_t(x_0 + x') - G_t(x_0) - J_{0\to t}(x_0)x'\big\|_2 \\
\le& \|G_t(x_0 + x') - G_t(x_0) - \widehat{G}_t^\delta(x_0 + x') + \widehat{G}_t^\delta(x_0)\|_2 + \|\widehat{G}_t^\delta(x_0 + x') - \widehat{G}_t^\delta(x_0) - J_{0\to t}^\delta(x_0)x'\|_2 \\
&+ \|J_{0\to t}^\delta(x_0)x' - J_{0\to t}(x_0)x'\|_2.
\end{aligned} \tag{61}$$

Note that

$$\|G_t(x_0 + x') - G_t(x_0) - \widehat{G}_t^\delta(x_0 + x') + \widehat{G}_t^\delta(x_0)\|_2 = \lim_{\delta'\to 0^+} \|\widehat{G}_t^{\delta'}(x_0 + x') - \widehat{G}_t^{\delta'}(x_0) - \widehat{G}_t^\delta(x_0 + x') + \widehat{G}_t^\delta(x_0)\|_2.$$

Without loss, we may only consider $x' \in \mathcal{B}(1)$. For all $\delta_1, \delta_2 \in (0, \delta]$, by the mean value theorem

$$\begin{aligned}
&\|\widehat{G}_t^{\delta_1}(x_0 + x') - \widehat{G}_t^{\delta_1}(x_0) - \widehat{G}_t^{\delta_2}(x_0 + x') + \widehat{G}_t^{\delta_2}(x_0)\|_2 \\
\le& \|\nabla\widehat{G}_t^{\delta_1}(x_0 + \alpha x') - \nabla\widehat{G}_t^{\delta_2}(x_0 + \alpha x')\|_{\mathrm{op}} \cdot \|x'\|_2 \\
\le& \limsup_{\delta_1,\delta_2\in(0,\delta]} \sup_{x\in\mathcal{B}(x_0,1)} \|J_{0\to t}^{\delta_1}(x) - J_{0\to t}^{\delta_2}(x)\|_{\mathrm{op}} \cdot \|x'\|_2 = c(\delta)\cdot\|x'\|_2,
\end{aligned}$$

where by Eq. (60) we have $\lim_{\delta\to 0^+} c(\delta) = 0$. Also by Eq. (60), we see that $\|J_{0\to t}^\delta(x_0)x' - J_{0\to t}(x_0)x'\|_2 \le c'(\delta)\cdot\|x'\|_2$, with $c'(\delta) \to 0^+$ as $\delta \to 0^+$. Plugging these results back into Eq. (61), we conclude that for any $\varepsilon > 0$, there exists $\delta_\varepsilon > 0$ such that for all $\delta \le \delta_\varepsilon$,

$$\big\|G_t(x_0 + x') - G_t(x_0) - J_{0\to t}(x_0)x'\big\|_2 \le \varepsilon\|x'\|_2 + \|\widehat{G}_t^\delta(x_0 + x') - \widehat{G}_t^\delta(x_0) - J_{0\to t}^\delta(x_0)x'\|_2.$$

By definition, we have $\lim_{\|x'\|_2\to 0^+} \|\widehat{G}_t^\delta(x_0 + x') - \widehat{G}_t^\delta(x_0) - J_{0\to t}^\delta(x_0)x'\|_2 = 0$. Since ε can be arbitrarily small, we then conclude that

$$\nabla_{x_0} G_t(x_0) = J_{0\to t}(x_0) \tag{62}$$

for all $t \in [0, T]$ and $x_0 \in \mathbb{R}^d$.

Finally, we are ready to prove the existence of a probability density. Recall that $\mathcal{I}_t(x_t) \subseteq \mathcal{B}(f_t(\|x_t\|_2))$. Therefore, for any $R > 0$ we have $G_t^{-1}(\mathcal{B}(R)) \subseteq \mathcal{B}(f_t(R))$. We choose $R \in \mathbb{R}_{>0}$ large enough such that $\|x_t\|_2 < R$. By Eq. (62) we know that the mappint $x_0 \mapsto x_t = G_t(x_0)$ has everywhere non-degenerate Jacobian matrix. Applying the inverse mapping theorem (Rudin et al., 1976), we conclude that for all $x \in \mathcal{B}(f_t(R))$, there exists an open set $\mathcal{S}(x)$ that contains x, such that G_t is injective on $\mathcal{S}(x)$, and the inverse is continuously differentiable. We denote this mapping by $h_{\mathcal{S}(x)}$ that is defined on $\mathcal{S}(x)$. By the Heine–Borel theorem (Borel, 1928), $\mathcal{B}(f_t(R))$ is covered by finitely many such $\mathcal{S}(x)$, and we denote by $\mathfrak{S}_R$ the collection of such $\mathcal{S}(x)$. As a consequence, we conclude that for all $x_t \in \mathbb{R}$, there are finitely many $x \in \mathbb{R}^d$ that satisfies $G_t(x) = x_t$, i.e., $|\mathcal{I}_t(x_t)| < \infty$. Therefore, $p_t(x_t) = \sum_{x\in\mathcal{I}_t(x_t)} p_0(x)\det(J_{0\to t}(x))^{-1}$. The proof is complete.

B.3 Proof of Theorem 4.2

We present in this section a rigorous proof of Theorem 4.2. Recall that we have proved in the first part of Appendix B.2 that $z_t = M_t z_0 + \xi_t$, where $M_t \in \mathbb{R}^{d \times d}$ and $\xi_t \in \mathbb{R}^d$ are functions of t only. We define $x_t' = M_t x_0 + \xi_t$, and denote its differential entropy by $H'(t)$. Through standard computation, we see that $H'(t)$ and $H_0(t)$ exist and satisfy

$$H'(t) = H(0) + \log\det(M_t) \leq H_0(0) + \log\det(M_t), \qquad H_0(t) = H_0(0) + \log\det(M_t).$$

Therefore, in order to show $H(t) \leq H_0(t)$, it suffices to prove $H(t) \leq H'(t)$. One caveat is that we still have to show $H(t)$ exists (in the sense of Lebesgue measure).

Recall that $\mathfrak{S}_R$ is a collection of covering sets introduced at the end of Appendix B.2. We can in fact choose the coverings appropriately such that for all $m \in \mathbb{N}_+$, $\mathfrak{S}_m \subseteq \mathfrak{S}_{m+1}$. Define $\mathfrak{S}_\infty = \cup_{m=1}^\infty \mathfrak{S}_m = \{\mathcal{S}_i : i \in \mathbb{N}_+\}$. Here, recall each $\mathcal{S}_i$ is an open set. For all $i \in \mathbb{N}_+$, we let $\bar{\mathcal{S}}_i = \mathcal{S}_i \backslash (\cup_{j=1}^{i-1} \mathcal{S}_j)$. Then it holds that $\bar{\mathcal{S}}_i \cap \bar{\mathcal{S}}_j = \varnothing$ for all $i \neq j$, and $\cup_{i=1}^\infty \bar{\mathcal{S}}_i = \mathbb{R}^d$. We denote by f_X the probability density function of a random variable X. Recall that in Appendix B.2 we have defined $x_t = G_t(x_0)$ and $\nabla_x G_t(x) = J_{0\to t}(x)$. Furthermore, by Eq. (60) it holds that $\det(J_{0\to t}(x)) \leq \det(M_t)$ for all $x \in \mathbb{R}^d$. Based on the derivations in Appendix B.2, we see that

$$f_{x_t}(x) = \sum_{z \in \mathcal{I}_t(x)} \sum_{i=1}^\infty \mathbb{1}\{z \in \bar{\mathcal{S}}_i\} f_{x_0}(z) \det(J_{0\to t}(z))^{-1}.$$

For $i \in \mathbb{N}_+$, we define $\bar{\mathcal{Z}}_i = \{x \in \mathbb{R}^d : M_t^{-1}(x - \xi_t) \in \bar{\mathcal{S}}_i\}$. Then $\bar{\mathcal{Z}}_i \cap \bar{\mathcal{Z}}_j = \varnothing$ for $i \neq j$, and $\cup_{i=1}^\infty \bar{\mathcal{Z}}_i = \mathbb{R}^d$. In addition,

$$\begin{aligned} -\int_{\bar{\mathcal{Z}}_i} f_{x_t'}(w) \log f_{x_t'}(w) \mathrm{d}w &= -\int_{\bar{\mathcal{S}}_i} f_{x_0}(z) \log\left[f_{x_0}(z) \det(M_t)^{-1}\right] \mathrm{d}z \\ &\geq -\int_{\bar{\mathcal{S}}_i} f_{x_0}(z) \log\left[f_{x_0}(z) \det(J_{0\to t}(z))^{-1}\right] \mathrm{d}z. \end{aligned} \tag{63}$$

By Eq. (60) we know that there exist constants $c_1, c_2 > 0$, such that $\det(J_{0\to t}(z)) \in (c_1, c_2)$ for all $z \in \mathbb{R}^d$. By assumption, the left hand side above has a finite Lebesgue integral, hence the Lebesgue integral in the second line of right hand side above also exists and is finite. Adding up the above terms over $i \in \mathbb{N}_+$ (recall that $h_{\mathcal{S}}$ is the restriction of G_t on S), we get

$$\begin{aligned} &-\int f_{x_t'}(w) \log f_{x_t'}(w) \mathrm{d}w \\ &\overset{(i)}{\geq} -\sum_{i=1}^\infty \int_{\bar{\mathcal{S}}_i} f_{x_0}(z) \log\left[f_{x_0}(z) \det(J_{0\to t}(z))^{-1}\right] \mathrm{d}z \\ &\overset{(ii)}{=} -\sum_{i=1}^\infty \int_{h_{\mathcal{S}_i}(\bar{\mathcal{S}}_i)} f_{x_0}(h_{S_i}^{-1}(x)) \det(J_{0\to t}(h_{S_i}^{-1}(x)))^{-1} \log\left[f_{x_0}(h_{S_i}^{-1}(x)) \det(J_{0\to t}(h_{S_i}^{-1}(x)))^{-1}\right] \mathrm{d}x \\ &\overset{(iii)}{=} -\int \sum_{i=1}^\infty \mathbb{1}_{h_{\mathcal{S}_i}(\bar{\mathcal{S}}_i)}(x) f_{x_0}(h_{S_i}^{-1}(x)) \det(J_{0\to t}(h_{S_i}^{-1}(x)))^{-1} \log\left[f_{x_0}(h_{S_i}^{-1}(x)) \det(J_{0\to t}(h_{S_i}^{-1}(x)))^{-1}\right] \mathrm{d}x \\ &\overset{(iv)}{\geq} -\int f_{x_t}(x) \log f_t(x) \mathrm{d}x. \end{aligned}$$

In the above display, the summation on the right hand side of (i) exists due to Lemma B.2, (ii) is by the change-of-variable technique for probability density functions, (iii) is also due to Lemma B.2, and (iv) is because

$$f_{x_t}(x) = \sum_{i=1}^\infty \mathbb{1}_{h_{\mathcal{S}_i}(\bar{\mathcal{S}}_i)}(x) f_{x_0}(h_{S_i}^{-1}(x)) \det(J_{0\to t}(h_{S_i}^{-1}(x)))^{-1}.$$

The proof is complete.

B.4 Technical lemmas

We collect in this section the technical lemmas that support proof in this section.

Lemma B.2. *We assume the conditions of Theorem 4.2. Then the following sum exists and is finite:*

$$-\sum_{i=1}^{\infty}\int_{\bar{\mathcal{S}}_i} f_{x_0}(z)\log\left[f_{x_0}(z)\det(J_{0\to t}(z))^{-1}\right]\mathrm{d}z.$$

Furthermore, we can exchange the order of integration and summation, in the sense that

$$\sum_{i=1}^{\infty}\int \mathbb{1}_{\bar{\mathcal{S}}_i}(z)f_{x_0}(z)\log\left[f_{x_0}(z)\det(J_{0\to t}(z))^{-1}\right]\mathrm{d}z=\int\sum_{i=1}^{\infty}\mathbb{1}_{\bar{\mathcal{S}}_i}(z)f_{x_0}(z)\log\left[f_{x_0}(z)\det(J_{0\to t}(z))^{-1}\right]\mathrm{d}z.$$

Proof of Lemma B.2. By Eq. (60), we know that there exist constants $c_1, c_2 > 0$, such that $\det(J_{0\to t}(z)) \in (c_1, c_2)$ for all $z \in \mathbb{R}^d$. Using this together with the assumption that the differential entropy of x_0 exists and is finite, we conclude that the function $z \mapsto f_{x_0}(z)\log\left[f_{x_0}(z)\det(J_{0\to t}(z))^{-1}\right]$ has a finite Lebesgue integral. The desired claims then immediately follow from the properties of Lebesgue integral. □

C Proofs related to the discretized process

C.1 Proof of Theorem 5.1

PROOF OF THE FIRST CLAIM

For all $k = 0, 1, \cdots, K-1$, by Eq. (18)

$$\begin{aligned}
&\langle X_{k+1}, \mu_y - \mu_{y'}\rangle - \langle X_k, \mu_y - \mu_{y'}\rangle\\
=&\,(e^{\delta_k}-1)e^{-(T-t_k)}\Big((1+\eta)\|\mu_y\|_2^2 - \eta\sum_{y''\in\mathcal{Y}} q_{T-t_k}(X_k, y'')\langle\mu_{y''}, \mu_y\rangle - (1+\eta)\langle\mu_y, \mu_{y'}\rangle\\
&+\eta\sum_{y''\in\mathcal{Y}} q_{T-t_k}(X_k, y'')\langle\mu_{y''}, \mu_{y'}\rangle\Big)\\
=&(e^{\delta_k}-1)e^{-(T-t_k)}\Big(\|\mu_y\|_2^2 - \langle\mu_y, \mu_{y'}\rangle + \eta\langle\mu_y - \mu_0, \mu_y - \mu_{y'}\rangle - \eta\sum_{y''\in\mathcal{Y}} q_{T-t_k}(X_k, y'')\langle\mu_{y''} - \mu_0, \mu_y - \mu_{y'}\rangle\Big)\Big)\\
\geq&(e^{\delta_k}-1)e^{-(T-t_k)}\Big(\|\mu_y\|_2^2 - \langle\mu_y, \mu_{y'}\rangle + \eta(1 - q_{T-t_k}(X_k, y))(\|\mu_y - \mu_0\|_2^2 - 3\varepsilon)\Big).
\end{aligned}\tag{64}$$

On the other hand, by definition

$$\langle Z_{k+1}, \mu_y - \mu_{y'}\rangle - \langle Z_k, \mu_y - \mu_{y'}\rangle = (e^{\delta_k}-1)e^{-(T-t_k)}\left(\|\mu_y\|_2^2 - \langle\mu_y, \mu_{y'}\rangle\right)$$

recall that we have assumed $\langle X_0, \mu_y - \mu_{y'}\rangle \geq \langle Z_0, \mu_y - \mu_{y'}\rangle$. By induction, we are able to conclude that $\langle X_k, \mu_y - \mu_{y'}\rangle \geq \langle Z_k, \mu_y - \mu_{y'}\rangle$ for all $y' \in \mathcal{Y}$ and $k \in \{0\} \cup [K]$. The first claim of the theorem then immediately follows.

PROOF OF THE SECOND CLAIM

The proof closely mirrors that of Theorem 3.6. Similar to the derivation of Eq. (28), we obtain that

$$\begin{aligned}
&\langle X_{k+1} - Z_{k+1}, \mu_y - \mu_{y'}\rangle - \langle X_k - Z_k, \mu_y - \mu_{y'}\rangle\\
\geq&(e^{\delta_k}-1)e^{-T+t_k}\left(\|\mu_y\|_2^2 - \langle\mu_y, \mu_{y'}\rangle + \eta e^{-\Delta/8}(\|\mu_y - \mu_0\|_2^2 - 3\varepsilon)\cdot\min\{1 - \mathcal{P}(X_k, y), \xi_w\}\right),
\end{aligned}$$

where we recall that $\xi_w = 1 - w_y/(w_y + \min_{y'\neq y} w_{y'})$. Now suppose $\langle X_k - Z_k, \mu_y - \mu_{y'}\rangle \in [0, \mathcal{U}]$ for all $k \in \{0\} \cup [K]$ and $y' \in \mathcal{Y}$, then like the derivation of Eq. (29), we get

$$\mathcal{P}(X_k, y) \leq 1 - \mathcal{F}(\max_{0\leq k\leq K}\mathcal{P}(Z_k, y), \mathcal{U})$$

for all $k \in \{0\} \cup [K]$. For $\mathcal{U}$ to be a valid upper bound, we must have

$$\mathcal{U} \geq\langle X_0 - Z_0, \mu_y - \mu_{y'}\rangle$$

$$
\begin{aligned}
&+\sum_{k=0}^{K-1}(e^{\delta_k}-1)e^{-T+t_k}\left(\|\mu_y\|_2^2-\langle\mu_y,\mu_{y'}\rangle+\eta e^{-\Delta/8}(\|\mu_y-\mu_0\|_2^2-3\varepsilon)\cdot\min\{\mathcal{F}(\max_{0\leq k\leq K}\mathcal{P}(Z_k,y),\mathcal{U}),\xi_w\}\right)\\
\geq&\langle X_0-Z_0,\mu_y-\mu_{y'}\rangle \qquad\qquad (65)\\
&+(e^{-T+t_K}-e^{-T})\left(\|\mu_y\|_2^2-\langle\mu_y,\mu_{y'}\rangle+\eta e^{-\Delta/8}(\|\mu_y-\mu_0\|_2^2-3\varepsilon)\cdot\min\{\mathcal{F}(\max_{0\leq k\leq K}\mathcal{P}(Z_k,y),\mathcal{U}),\xi_w\}\right).
\end{aligned}
$$

For any $\mathcal{U}$ that does not satisfy Eq. (65), we know that $\langle X_K-Z_K,\mu_y-\mu_{y'}\rangle\geq\mathcal{U}$, and

$$
\mathcal{P}(X_K,y)\geq\frac{\mathcal{P}(Z_K,y)}{\mathcal{P}(Z_K,y)+(1-\mathcal{P}(Z_K,y))\cdot\exp(-\mathcal{U})}, \tag{66}
$$

completing the proof of the first result. As for the proof of the convergence rate, we simply set $e^{-\mathcal{U}}=\eta^{-1}(\log\eta)^2$. For such $\mathcal{U}$, the left hand side of Eq. (65) is of order $O(\log\eta)$, while the right hand side of Eq. (65) is of order $O(\eta\wedge(\log\eta)^2)$. We then conclude that for a large enough η, Eq. (65) does not hold. Plugging such $\mathcal{U}$ back into Eq. (66), we are able to deduce the desired convergence rate.

C.2 Proof of Theorem 5.2

For $k\in\{0\}\cup[K-1]$, we define

$$
F_k(x,\eta)=x+(e^{\delta_k}-1)\cdot\Big(x-\Sigma_{T-t_k}^{-1}x+e^{-T+t_k}(1+\eta)\Sigma_{T-t_k}^{-1}\mu_y-\eta e^{-T+t_k}\sum_{y'\in\mathcal{Y}}q_{T-t_k}(x,y')\Sigma_{T-t_k}^{-1}\mu_{y'}\Big).
$$

Observe that $X_{k+1}=F_k(X_k,\eta)$ and $Z_{k+1}=F_k(Z_k,0)$. We then take the gradient of F_k with respect to the first argument, which gives

$$
\nabla_x F_k(x,\eta)=I_d+(e^{\delta_k}-1)\Big(I_d-\Sigma_{T-t_k}^{-1}-\eta e^{-T+t_k}\Omega_{T-t_k}(x)\Big),
$$

where

$$
\begin{aligned}
&\Omega_{T-t_k}(x)\\
=&\sum_{y'\in\mathcal{Y}}q_{T-t_k}(x,y')\Sigma_{T-t_k}^{-1}\mu_{y'}\mu_{y'}^\top\Sigma_{T-t_k}^{-1}-\Big(\sum_{y'\in\mathcal{Y}}q_{T-t_k}(x,y')\Sigma_{T-t_k}^{-1}\mu_{y'}\Big)\Big(\sum_{y'\in\mathcal{Y}}q_{T-t_k}(x,y')\Sigma_{T-t_k}^{-1}\mu_{y'}\Big)^\top.
\end{aligned}
$$

We then conclude that $\nabla_x F_k(x,\eta)\preceq\nabla_x F_k(x,0)$ for all $\eta\geq 0$. We denote by $\sigma_{\min}(X)$ the minimum eigenvalue of a matrix X. Observe that

$$
\begin{aligned}
\sigma_{\min}(\nabla_x F_k(x,\eta))&\geq e^{\delta_k}-\frac{e^{\delta_k}-1}{\sigma_{\min}(\Sigma)\wedge 1}-\frac{(e^{\delta_k}-1)\eta\sup_{y'\in\mathcal{Y}}\|\mu_{y'}\|_2^2}{\sigma_{\min}(\Sigma)^2\wedge 1},\\
\sigma_{\min}(\nabla_x F_k(x,0))&\geq e^{\delta_k}-\frac{e^{\delta_k}-1}{\sigma_{\min}(\Sigma)\wedge 1},
\end{aligned} \tag{67}
$$

both are strictly positive under the current set of assumptions. Observe that $x\mapsto F_k(x,0)$ is an affine transformation, then it is also bijective, while $F_k(x,\eta)$ is not necessarily one-to-one. For $x\in\mathbb{R}^d$ and $\eta\geq 0$, we let $G_k(x,\eta):=F_k(F_{k-1}(\cdots F_0(x,\eta)\cdots,\eta),\eta)$. Then $X_{k+1}=G_k(X_0,\eta)$ and $Z_{k+1}=G_k(Z_0,0)$. Observe that there exists $A_k\in\mathbb{R}^{d\times d}$ and $\beta_k\in\mathbb{R}^d$, such that $G_k(x,0)=A_kx+\beta_k$. In addition, one can verify that A_k is non-degenerate.

In the sequel, we use f_X to represent the probability density function of a random variable X. Utilizing a change-of-variable technique, we see that $\det(A_k)\cdot f_{Z_{k+1}}(A_kx+\beta_k)=f_{Z_0}(x)$ for all $x\in\mathbb{R}^d$. We can then express the differential entropy of Z_{k+1} based on that of Z_0:

$$
\mathcal{H}_0(k+1)=-\int f_{Z_{k+1}}(x)\log f_{Z_{k+1}}(x)\mathrm{d}x=\mathcal{H}_0(0)+\log\det(A_k).
$$

In the next lemma, we demonstrate the existence of probability density function of X_{k+1} with respect to the Lebesgue measure.

Lemma C.1. *We assume the conditions of Theorem 5.2. Then, for all $k \in \{0\} \cup [K]$, X_k has a probability density function with with respect to the Lebesgue measure.*

The proof of Lemma C.1 is similar to that of Lemma B.1, and we skip it for the compactness of presentation. We define $X'_{k+1} = A_k X_0 + \beta_k$, and denote the differential entropy of X'_{k+1} by $\mathcal{H}'(k+1)$. Similarly, we have $\mathcal{H}'(k+1) = \mathcal{H}(0) + \log\det(A_k) \leq \mathcal{H}_0(0) + \log\det(A_k)$. Therefore, in order to prove $\mathcal{H}(k+1) \leq \mathcal{H}_0(k+1)$, it suffices to show $\mathcal{H}(k+1) \leq \mathcal{H}'(k+1)$.

The remainder proof follows analogously as that of Lemma B.1. Taking the Jacobian matrix of $G_k(x, \eta)$ with respect to the first argument, we get

$$\nabla G_k(x_0, \eta) = \nabla F_k(x_k, \eta) \cdot \nabla F_{k-1}(x_{k-1}, \eta) \cdots \nabla F_0(x_0, \eta),$$

where $x_i = G_{i-1}(x_0, \eta)$. By Eq. (67), we know that $\nabla G_k(x_0, \eta)$ is everywhere non-degenerate. In addition, by induction we conclude that $\|X_{k+1}\|_2 \geq \alpha_{k+1}\|X_0\|_2 - \gamma_{k+1}$, where $\alpha_{k+1} \in \mathbb{R}_{>0}$ and $\gamma_k \in \mathbb{R}$. We define

$$\mathcal{I}_{k+1}(X_{k+1}) := \{X_0 \in \mathbb{R}^d, G_k(X_0, \eta) = X_{k+1}\}.$$

Then $\mathcal{I}_{k+1}(X_{k+1}) \subseteq \mathcal{B}(\alpha_{k+1}^{-1}(\|X_{k+1}\|_2 + \gamma_{k+1}))$. By the inverse mapping theorem, we obtain that for all $x \in \mathbb{R}^d$, there exists an open set $\mathcal{S}(x)$ that contains x, such that $G_k(\cdot, \eta)$ is injective on $\mathcal{S}(x)$. We denote this injection by $h_{\mathcal{S}}$. By Heine–Borel theorem, $\mathcal{B}(\alpha_{k+1}^{-1}(\|X_{k+1}\|_2 + \gamma_{k+1}))$ can be covered by finitely many $\mathcal{S}(x)$. Therefore, for all $x \in \mathbb{R}^d$, we have $f_{X_{k+1}}(x) = \sum_{z \in \mathcal{I}_{k+1}(x)} f_{X_0}(z) \det(\nabla G_k(z, \eta))^{-1}$. In addition, $f_{X'_{k+1}}(A_k z + \beta_k) = f_{X_0}(z) \cdot \det(A_k)^{-1}$ and $\det(A_k)^{-1} \leq \det(\nabla G_k(z, \eta))^{-1}$. By the assumption that the differential entropy of X_0 exists and is finite, we may also conclude that the differential entropy of X'_{k+1} exists and is finite.

When $\alpha_{k+1}^{-1}(\|X_{k+1}\|_2 + \gamma_{k+1}) = R$, we denote by $\mathfrak{S}_R$ the collection of such $\mathcal{S}(x)$. We can construct $\mathfrak{S}_R$ for every positive R. It is not hard to see that we can choose the coverings appropriately such that for all $m \in \mathbb{N}_+$, we have $\mathfrak{S}_m \subseteq \mathfrak{S}_{m+1}$. Consider the union of these covering sets: $\mathfrak{S}_\infty = \cup_{m=1}^\infty \mathfrak{S}_m = \{\mathcal{S}_i : i \in \mathbb{N}_+\}$. We define $\bar{\mathcal{S}}_i = \mathcal{S}_i \backslash (\cup_{j=1}^{i-1} \mathcal{S}_j)$, and let $\bar{\mathcal{Z}}_i = \{x \in \mathbb{R}^d : A_k^{-1}(x - \beta_k) \in \bar{\mathcal{S}}_i\}$. Note that $\bar{\mathcal{S}}_i \cap \bar{\mathcal{S}}_j = \varnothing$ for all $i \neq j$ and $\cup_{i=1}^\infty \bar{\mathcal{S}}_i = \mathbb{R}^d$. Then, it hold that

$$f_{X_{k+1}}(x) = \sum_{z \in \mathcal{I}_{k+1}(x)} \sum_{i=1}^{\infty} \mathbb{1}\{z \in \bar{\mathcal{S}}_i\} f_{X_0}(z) \det(\nabla G_k(z, \eta))^{-1}$$

Note that for all $i \in \mathbb{N}_+$,

$$\begin{aligned} -\int_{\bar{\mathcal{Z}}_i} f_{X'_{k+1}}(w) \log f_{X'_{k+1}}(w) \mathrm{d}x &= -\int_{\bar{\mathcal{S}}_i} f_{X_0}(z) \log\left[f_{X_0}(z) \cdot \det(A_k)^{-1}\right] \mathrm{d}z \\ &\geq -\int_{\bar{\mathcal{S}}_i} f_{X_0}(z) \log\left[f_{X_0}(z) \cdot \det(\nabla G_k(z, \eta))^{-1}\right] \mathrm{d}z. \end{aligned}$$

Summing both sides of the above equality over i, we get

$$\begin{aligned} & -\int f_{X'_{k+1}}(w) \log f_{X'_{k+1}}(w) \mathrm{d}w \\ & \geq -\sum_{i \in \mathbb{N}_+} \int_{\bar{\mathcal{S}}_i} f_{X_0}(z) \log\left[f_{X_0}(z) \cdot \det(\nabla G_k(z, \eta))^{-1}\right] \mathrm{d}z. \\ & = -\sum_{i \in \mathbb{N}_+} \int_{h_{\mathcal{S}_i}(\bar{\mathcal{S}}_i)} f_{X_0}(h_{\mathcal{S}_i}^{-1}(x)) \cdot \det(\nabla G_k(h_{\mathcal{S}_i}^{-1}(x), \eta))^{-1} \log\left[f_{X_0}(h_{\mathcal{S}_i}^{-1}(x)) \cdot \det(\nabla G_k(h_{\mathcal{S}_i}^{-1}(x), \eta))^{-1}\right] \mathrm{d}x \\ & \overset{(i)}{=} -\int \sum_{i \in \mathbb{N}_+} \mathbb{1}\{x \in h_{\mathcal{S}_i}(\bar{\mathcal{S}}_i)\} f_{X_0}(h_{\mathcal{S}_i}^{-1}(x)) \cdot \det(\nabla G_k(h_{\mathcal{S}_i}^{-1}(x), \eta))^{-1} \log\left[f_{X_0}(h_{\mathcal{S}_i}^{-1}(x)) \cdot \det(\nabla G_k(h_{\mathcal{S}_i}^{-1}(x), \eta))^{-1}\right] \mathrm{d}x \\ & \geq -\int f_{X_{k+1}}(x) \log f_{X_{k+1}}(x) \mathrm{d}x, \end{aligned}$$

where to exchange the order of summation and integration in (i) we make use of the assumption that the differential entropy of X_0 exists and is finite. The proof is complete.

C.3 Results for the DDPM sampler

As for the DDPM sampler, we can only establish results for classification confidence. The proof is deferred to Appendix C.4.

Theorem C.2. *We assume the conditions of Theorem 5.1. Then we have the following results:*

1. *For all $k \in \{0\} \cup [K]$, it holds that $\mathcal{P}(\bar{X}_k, y) \geq \mathcal{P}(\bar{Z}_k, y)$.*
2. *If additionally we assume $\Delta_{\max} = \max_{k \in \{0\} \cup [K-1]} \delta_k \leq 1/2$, then*

$$\mathcal{P}(\bar{X}_K, y) \geq \frac{\mathcal{P}(\bar{Z}_K, y)}{\mathcal{P}(\bar{Z}_K, y) + (1 - \mathcal{P}(\bar{Z}_K, y)) \cdot \exp(-e^{-3T}\bar{\mathcal{U}})} \geq \mathcal{P}(\bar{Z}_K, y),$$

where $\bar{\mathcal{U}} \in \mathbb{R}_{\geq 0}$ is any number that satisfies

$$\begin{aligned}&\bar{\mathcal{U}} - e^{-T-\Delta_{\max}}\langle \bar{X}_0 - \bar{Z}_0, \mu_y - \mu_{y'}\rangle \\ &< \eta e^{-\Delta/8}(e^{-2T} - e^{-4T}) \min\{\mathcal{F}\big(\max_{0 \leq k \leq K} \mathcal{P}(\bar{Z}_k, y), \bar{\mathcal{U}}\big), \xi_w\}(\|\mu_y - \mu_0\|_2^2 - 3\varepsilon).\end{aligned}$$

We recall that $(\mathcal{F}, \Delta, \xi_w)$ are defined in Eq. (16). In addition, as $\eta \to \infty$, the convergence rate is at least $1 - O(\eta^{-e^{-T}}(\log \eta)^{2e^{-T}})$.

Remark C.3. We can eliminate the assumptions on the component centers for Theorem 5.1 and C.2 when $|\mathcal{Y}| = 2$. The proof is similar to that of Theorem 3.8 and 3.10, and we skip it for the sake of simplicity.

C.4 Proof of Theorem C.2

PROOF OF THE FIRST CLAIM

Observe that for the guided process,

$$\begin{aligned}&\langle \bar{X}_{k+1}, \mu_y - \mu_{y'}\rangle \\ =&(2 - e^{\delta_k})\langle \bar{X}_k, \mu_y - \mu_{y'}\rangle + 2(e^{\delta_k} - 1)e^{-(T-t_k)}\Big((1+\eta)\|\mu_y\|_2^2 - \eta \sum_{y'' \in \mathcal{Y}} q_{T-t_k}(\bar{X}_k, y'')\langle \mu_{y''}, \mu_y\rangle - (1+\eta)\langle \mu_y, \mu_{y'}\rangle \\ &+ \eta \sum_{y'' \in \mathcal{Y}} q_{T-t_k}(\bar{X}_k, y'')\langle \mu_{y''}, \mu_{y'}\rangle\Big) + \sqrt{2(e^{\delta_k} - 1)}\langle W_k, \mu_y - \mu_{y'}\rangle \\ \geq&(2 - e^{\delta_k})\langle \bar{X}_k, \mu_y - \mu_{y'}\rangle + 2(e^{\delta_k} - 1)e^{-(T-t_k)}\Big(\|\mu_y\|_2^2 - \langle \mu_y, \mu_{y'}\rangle + \eta(1 - q_{T-t_k}(\bar{X}_k, y))(\|\mu_y - \mu_0\|_2^2 - 3\varepsilon)\Big) \\ &+ \sqrt{2(e^{\delta_k} - 1)}\langle W_k, \mu_y - \mu_{y'}\rangle.\end{aligned}$$

As for the unguided process, we have

$$\langle \bar{Z}_{k+1}, \mu_y - \mu_{y'}\rangle = (2 - e^{\delta_k})\langle \bar{Z}_k, \mu_y - \mu_{y'}\rangle + 2(e^{\delta_k} - 1)e^{-(T-t_k)}(\|\mu_y\|_2^2 - \langle \mu_y, \mu_{y'}\rangle) + \sqrt{2(e^{\delta_k} - 1)}\langle W_k, \mu_y - \mu_{y'}\rangle.$$

By induction, we know that $\langle \bar{X}_k, \mu_y - \mu_{y'}\rangle \geq \langle \bar{Z}_k, \mu_y - \mu_{y'}\rangle$ for all $k \in \{0\} \cup [K]$. The first claim then immediately follows.

PROOF OF THE SECOND CLAIM

Similar to the derivation of Eq. (37), we obtain that

$$\begin{aligned}&\langle \bar{X}_{k+1} - \bar{X}_k, \mu_y - \mu_{y'}\rangle \\ \geq&2(e^{\delta_k} - 1)e^{-T+t_k}\Big(\|\mu_y\|_2^2 - \langle \mu_y, \mu_{y'}\rangle + \eta e^{-\Delta/8} \min\{1 - \mathcal{P}(\bar{X}_k, y), \xi_w\}(\|\mu_y - \mu_0\|_2^2 - 3\varepsilon)\Big) + \\ &- (e^{\delta_k} - 1)\langle \bar{X}_k, \mu_y - \mu_{y'}\rangle + \sqrt{2(e^{\delta_k} - 1)}\langle W_k, \mu_y - \mu_{y'}\rangle.\end{aligned}$$

As for the unguided process, we have

$$\begin{aligned}\langle \bar{Z}_{k+1} - \bar{Z}_k, \mu_y - \mu_{y'} \rangle = &-(e^{\delta_k} - 1)\langle \bar{Z}_k, \mu_y - \mu_{y'} \rangle + 2(e^{\delta_k} - 1)e^{-T+t_k}\Big(\|\mu_y\|_2^2 - \langle \mu_y, \mu_{y'} \rangle\Big) \\ &+ \sqrt{2(e^{\delta_k} - 1)}\langle W_k, \mu_y - \mu_{y'} \rangle.\end{aligned}$$

Taking the difference, we see that

$$\begin{aligned}&\langle \bar{X}_{k+1} - \bar{Z}_{k+1}, \mu_y - \mu_{y'} \rangle \\ \geq &(2 - e^{\delta_k})\langle \bar{X}_k - \bar{Z}_k, \mu_y - \mu_{y'} \rangle + 2(e^{\delta_k} - 1)\eta e^{-T+t_k-\Delta/8} \min\{1 - \mathcal{P}(\bar{X}_k, y), \xi_w\}(\|\mu_y - \mu_0\|_2^2 - 3\varepsilon).\end{aligned} \tag{68}$$

From the above equation as well as our initial assumption we know that $\langle \bar{X}_k - \bar{Z}_k, \mu_y - \mu_{y'} \rangle \geq 0$ for all $k \in \{0\} \cup [K]$. Now suppose $\langle \bar{X}_k - \bar{Z}_k, \mu_y - \mu_{y'} \rangle \in [0, \bar{\mathcal{U}}]$ for all $k \in \{0\} \cup [K]$. Then similar to the derivation of Eq. (29), we know that

$$\mathcal{P}(\bar{X}_k, y) \leq 1 - \mathcal{F}\big(\max_{0 \leq k \leq K} \mathcal{P}(\bar{Z}_k, y), \bar{\mathcal{U}}\big) \tag{69}$$

for all $k \in \{0\} \cup [K]$. By Eq. (68) and (69) and induction hypothesis, we get the following lower bound:

$$\begin{aligned}&\langle \bar{X}_K - \bar{Z}_K, \mu_y - \mu_{y'} \rangle - e^{-3T}\langle \bar{X}_0 - \bar{Z}_0, \mu_y - \mu_{y'} \rangle \\ \geq &\sum_{k=0}^{K-1} 2\eta(e^{\delta_k} - 1)e^{-T+t_k-\Delta/8} \min\{\mathcal{F}\big(\max_{0 \leq k \leq K} \mathcal{P}(\bar{Z}_k, y), \bar{\mathcal{U}}\big), \xi_w\}(\|\mu_y - \mu_0\|_2^2 - 3\varepsilon) \cdot \prod_{j=k+1}^{K-1}(2 - e^{\delta_j}) \\ \geq &\sum_{k=0}^{K-1} 2\eta\delta_k e^{-T+2t_{k+1}-\Delta/8} \min\{\mathcal{F}\big(\max_{0 \leq k \leq K} \mathcal{P}(\bar{Z}_k, y), \bar{\mathcal{U}}\big), \xi_w\}(\|\mu_y - \mu_0\|_2^2 - 3\varepsilon) \cdot e^{-3T} \\ \geq &\eta e^{-\Delta/8}(e^{-2T} - e^{-4T}) \min\{\mathcal{F}\big(\max_{0 \leq k \leq K} \mathcal{P}(\bar{Z}_k, y), \bar{\mathcal{U}}\big), \xi_w\}(\|\mu_y - \mu_0\|_2^2 - 3\varepsilon).\end{aligned} \tag{70}$$

Then for $\bar{\mathcal{U}}$ to serve as a valid upper bound, we must have

$$\begin{aligned}&\bar{\mathcal{U}} - e^{-3T}\langle \bar{X}_0 - \bar{Z}_0, \mu_y - \mu_{y'} \rangle \\ \geq &\eta e^{-\Delta/8}(e^{-2T} - e^{-4T}) \min\{\mathcal{F}\big(\max_{0 \leq k \leq K} \mathcal{P}(\bar{Z}_k, y), \bar{\mathcal{U}}\big), \xi_w\}(\|\mu_y - \mu_0\|_2^2 - 3\varepsilon).\end{aligned} \tag{71}$$

Hence, if Eq. (71) is not satisfied, then there exists $k \in \{0\} \cup [K]$ such that $\langle \bar{X}_k - \bar{Z}_k, \mu_y - \mu_{y'} \rangle \geq \bar{\mathcal{U}}$. As a consequence, we have $\langle \bar{X}_K - \bar{Z}_K, \mu_y - \mu_{y'} \rangle \geq \prod_{j=k}^{K-1}(2 - e^{\delta_j})\bar{\mathcal{U}} \geq e^{-3T}\bar{\mathcal{U}}$, which further implies that

$$\mathcal{P}(\bar{X}_K, y) \geq \frac{\mathcal{P}(\bar{Z}_K, y)}{\mathcal{P}(\bar{Z}_K, y) + (1 - \mathcal{P}(\bar{Z}_K, y)) \cdot \exp(-e^{-3T}\bar{\mathcal{U}})}.$$

The proof of the first result is complete. The proof of the convergence rate follows analogously as that of the second part of Theorems 3.7 and 3.10. Here, we skip it for the compactness of presentation.

C.5 Proofs in Section 5.2

Assumption 3.1 does not hold for μ_{neg} It suffices to argue for the center component in μ_{neg}. Suppose for contradiction that there exists a vector μ_0 and a positive ε satisfying

$$|\langle -\mu_0, \mu - \mu_0 \rangle| \leq \varepsilon, \quad |\langle -\mu_0, -\mu - \mu_0 \rangle| \leq \varepsilon, \quad \|\mu_0\|_2^2/3 \geq \varepsilon.$$

Rewriting the first two inequalities, we obtain

$$\big|\langle \mu_0, \mu \rangle + \|\mu_0\|_2^2\big| \leq \varepsilon \quad \text{and} \quad \big|- \langle \mu_0, \mu \rangle + \|\mu_0\|_2^2\big| \leq \varepsilon.$$

Due to the symmetry, we assume without loss of generality that $\langle \mu_0, \mu \rangle \geq 0$. By comparing

$$\langle \mu_0, \mu \rangle + \|\mu_0\|_2^2 \leq \varepsilon \quad \text{with} \quad \|\mu_0\|_2^2/3 \geq \varepsilon,$$

we must have $\mu_0 = 0$ and $\varepsilon = 0$. This contradicts the fact that ε is positive. Therefore, the first item in Assumption 3.1 does not hold.

Proof of DDIM We focus on generating the center component $\mathsf{N}(0, I_d)$. Setting the guidance strength parameter η and using the discretized DDIM backward process yield

$$X_{k+1} = e^{\delta_k} X_k - (e^{\delta_k} - 1)\eta e^{-T+t_k} \frac{\exp(e^{-T+t_k} X_k^\top \mu) - \exp(-e^{-T+t_k} X_k^\top \mu)}{\exp(e^{-2T+2t_k}\|\mu\|_2^2/2) + \exp(e^{-T+t_k} X_k^\top \mu) + \exp(-e^{-T+t_k} X_k^\top \mu)} \mu. \tag{72}$$

Taking inner product with μ on both sides of Eq. (72) gives rise to

$$\begin{aligned}
&\langle X_{k+1}, \mu\rangle - \langle X_k, \mu\rangle \\
&= (e^{\delta_k} - 1)\langle X_k, \mu\rangle - (e^{\delta_k} - 1)\eta e^{-T+t_k} \frac{\exp(e^{-T+t_k} X_k^\top \mu) - \exp(-e^{-T+t_k} X_k^\top \mu)}{\exp(e^{-2T+2t_k}\|\mu\|_2^2/2) + \exp(e^{-T+t_k} X_k^\top \mu) + \exp(-e^{-T+t_k} X_k^\top \mu)} \|\mu\|_2^2.
\end{aligned}$$

We denote $v_k = \langle X_k, \mu\rangle$ and cast the last display into

$$v_{k+1} - v_k = (e^{\delta_k} - 1)v_k - (e^{\delta_k} - 1)\eta e^{-T+t_k} \frac{\exp(e^{-T+t_k} v_k) - \exp(-e^{-T+t_k} v_k)}{\exp(e^{-2T+2t_k}\|\mu\|_2^2/2) + \exp(e^{-T+t_k} v_k) + \exp(-e^{-T+t_k} v_k)} \|\mu\|_2^2. \tag{73}$$

We show the following stronger version of Proposition 5.3.

Lemma C.4. *Consider the Gaussian mixture model in Eq. (20). There exist positive constants $\eta_0 \le \frac{1}{\|\mu\|_2^2(\max e^{\delta_k} - 1)}$ and η_0' that depend on discretization step sizes $\{\delta_k\}_{k=0}^{K-1}$, such that for any k verifying $e^{-T+t_k} \ge 1/2$, it holds that*

1. *when $\eta \le \eta_0$, v_k evolves towards 0, i.e., $|v_{k+1}| < |v_k|$ if $v_k \neq 0$. Furthermore, for a small $\gamma \in (0, 1)$ satisfying*

$$\frac{(\gamma + e^{\delta_k} - 1)(\exp(\|\mu\|^2/2) + 2\exp(|v_k|))^2}{2\exp(\frac{\|\mu\|^2}{8}) + 4} \le (e^{\delta_k} - 1)\eta e^{-2T+2t_k}\|\mu\|_2^2 \le \frac{(2 - 2\gamma + 2e^{\delta_k})\left(2 + \exp\left(\frac{\|\mu\|_2^2}{8}\right)\right)^2}{8 + \left(2 + \exp\left(\frac{\|\mu\|_2^2}{8}\right)\right)^2},$$

   we have $|v_{k+1}| \le (1 - \gamma)|v_k|$. One can verify the existence of such a γ when δ_k is sufficiently small.

2. *when $\eta \ge \eta_0'$, there exists positive a and b dependent on η, and it holds that*

$$\begin{aligned}
|v_{k+1}| > |v_k| \quad &\textit{if} \quad |v_k| \in (0, a]; \\
|v_{k+1}| < |v_k| \quad &\textit{if} \quad |v_k| > b.
\end{aligned}$$

   In particular, thresholds a and b increase as η increases.

Proof. We first discuss the case when $v_k \ge 0$ and study the solution of the equation

$$2v_k = -(e^{\delta_k} - 1)v_k + (e^{\delta_k} - 1)\eta e^{-T+t_k} \frac{\exp(e^{-T+t_k} v_k) - \exp(-e^{-T+t_k} v_k)}{\exp(e^{-2T+2t_k}\|\mu\|_2^2/2) + \exp(e^{-T+t_k} v_k) + \exp(-e^{-T+t_k} v_k)} \|\mu\|_2^2. \tag{74}$$

Intuitively, the solution of Eq. (74) implies that for such a v_k, after one iteration, it holds that $v_{k+1} = -v_k$. To simplify the notation, we denote $\iota_k = e^{\delta_k} - 1$ and

$$h(v_k, k) = -\iota_k v_k + \iota_k \eta e^{-T+t_k} \frac{\exp(e^{-T+t_k} v_k) - \exp(-e^{-T+t_k} v_k)}{\exp(e^{-2T+2t_k}\|\mu\|_2^2/2) + \exp(e^{-T+t_k} v_k) + \exp(-e^{-T+t_k} v_k)} \|\mu\|_2^2. \tag{75}$$

To prove the lemma, below we will establish the following dichotomy for appropriate η_0 and η_0': 1) when $\eta < \eta_0$, $v_k = 0$ is the only solution to $h(v_k, k) = 0$; 2) when $\eta > \eta_0'$, $h(v_k, k) = 0$ has multiple solutions.

Proof of the first claim Taking the derivative of $h(v_k,k)$ with respect to v_k gives

$$\frac{\partial h}{\partial v_k}(v_k,k) = -\iota_k + \iota_k \eta e^{-2T+2t_k}\|\mu\|_2^2 \frac{\exp(e^{-2T+2t_k}\|\mu\|_2^2/2)[\exp(e^{-T+t_k}v_k)+\exp(-e^{-T+t_k}v_k)]+4}{[\exp(e^{-2T+2t_k}\|\mu\|_2^2/2)+\exp(e^{-T+t_k}v_k)+\exp(-e^{-T+t_k}v_k)]^2} - 2. \tag{76}$$

We then choose η_0 small enough, such that $\iota_k \eta e^{-2T+2t_k}\|\mu\|_2^2 \leq 1$ for all $\eta \leq \eta_0$, which allows us to set $\eta_0 \leq \frac{1}{\|\mu\|_2^2 \max \iota_k}$. In this case, Eq. (76) is always negative for any $v_k \geq 0$ and k. To see this, note that

$$\begin{aligned}
\frac{\partial h}{\partial v_k}(v_k,k) &\leq \frac{\exp(e^{-2T+2t_k}\|\mu\|_2^2/2)[\exp(e^{-T+t_k}v_k)+\exp(-e^{-T+t_k}v_k)]+4}{[\exp(e^{-2T+2t_k}\|\mu\|_2^2/2)+\exp(e^{-T+t_k}v_k)+\exp(-e^{-T+t_k}v_k)]^2} - 2 - \iota_k \\
&= -\iota_k - \frac{2\exp(e^{-2T+2t_k}\|\mu\|_2^2)+2\exp(2e^{-T+t_k}v_k)+2\exp(-2e^{-T+t_k}v_k)}{[\exp(e^{-2T+2t_k}\|\mu\|_2^2/2)+\exp(e^{-T+t_k}v_k)+\exp(-e^{-T+t_k}v_k)]^2} \\
&\quad - \frac{3\exp(e^{-2T+2t_k}\|\mu\|_2^2/2)[\exp(e^{-T+t_k}v_k)+\exp(-e^{-T+t_k}v_k)]}{[\exp(e^{-2T+2t_k}\|\mu\|_2^2/2)+\exp(e^{-T+t_k}v_k)+\exp(-e^{-T+t_k}v_k)]^2} \\
&< 0.
\end{aligned}$$

As a consequence, $h(v_k,k)$ is strictly decreasing for $v_k \geq 0$ as demonstrated in the left panel of Figure 4. It is straightforward to check that $h(0,k)=0$. Therefore, 0 is the only solution to $h(v_k,k)=0$. We define $\widetilde{h}(v_k,k) = h(v_k,k)+2v_k$. Then $\widetilde{h}(v_k,k) < 2v_k$ for all $v_k > 0$.

For the case when $v_k \leq 0$, By symmetry, we have $\widetilde{h}(v_k,k) = -\widetilde{h}(-v_k,k)$ and therefore, for $v_k > 0$, it holds that $\widetilde{h}(v_k,k) = -\widetilde{h}(-v_k,k) > 2v_k$ given $\eta_0 \leq \frac{1}{\|\mu\|_2^2 \max \iota_k}$. As a result, we deduce that $\widetilde{h}(v_k,k) < 2|v_k|$ for all $v_k \neq 0$. Observe that $\widetilde{h}(v_k,k) > 0$ for $v_k > 0$, and $\widetilde{h}(v_k,k) < 0$ for all $v_k < 0$. Therefore, rewriting Eq. (73), we get

$$|v_{k+1}| = |\widetilde{h}(v_k,k) - v_k| < |v_k| \quad \text{for} \quad v_k \neq 0.$$

Setting $\eta_0 = \frac{1}{\|\mu\|_2^2 \max \iota_k}$ proves the claim that $|v_{k+1}| < |v_k|$ if $v_k \neq 0$.

We can further show that when $\iota_k \eta e^{-2T+2t_k}\|\mu\|_2^2$ is sufficiently small, we can guarantee a strict magnitude shrinkage of $|v_k|$, i.e., $|v_{k+1}| \leq (1-\gamma)|v_k|$ for some small $\gamma \in (0,1)$. To see this, we aim to show a sandwich inequality when v_k is non-negative:

$$\gamma v_k \leq \widetilde{h}(v_k,k) \leq (2-\gamma)v_k.$$

Accordingly, we denote

$$\begin{aligned}
h_{2-\gamma}(v_k,k) &= \iota_k \eta e^{-T+t_k} \frac{\exp(e^{-T+t_k}v_k) - \exp(-e^{-T+t_k}v_k)}{\exp(e^{-2T+2t_k}\|\mu\|_2^2/2)+\exp(e^{-T+t_k}v_k)+\exp(-e^{-T+t_k}v_k)}\|\mu\|_2^2 - (2-\gamma+\iota_k)v_k, \\
h_{\gamma}(v_k,k) &= \iota_k \eta e^{-T+t_k} \frac{\exp(e^{-T+t_k}v_k) - \exp(-e^{-T+t_k}v_k)}{\exp(e^{-2T+2t_k}\|\mu\|_2^2/2)+\exp(e^{-T+t_k}v_k)+\exp(-e^{-T+t_k}v_k)}\|\mu\|_2^2 - (\gamma+\iota_k)v_k.
\end{aligned} \tag{77}$$

It is obvious that $v_k = 0$ is a zero point of the two functions stated in Eq. (77). To show that $|v_{k+1}| \leq (1-\gamma)|v_k|$, we need to show that for a sufficiently small $\iota_k \eta e^{-2T+2t_k}\|\mu\|_2^2$, $h_{2-\gamma} \leq 0$ and $h_\gamma \geq 0$ for all $v_k \geq 0$. We adopt the notations $a_k = \exp(e^{-2T+2t_k}\|\mu\|_2^2/2)$, $b_k = \exp(e^{-T+t_k}v_k)$ and $m_k = \iota_k \eta e^{-2T+2t_k}\|\mu\|_2^2$. To ensure $h_{2-\gamma}(v_k,k) \leq 0$, it suffices to find a sufficiently small m_k, such that

$$m_k \cdot \frac{a_k(b_k+1/b_k)+4}{(a_k+b_k+1/b_k)^2} \leq 2-\gamma+\iota_k. \tag{78}$$

Since $xy \leq (x+y)^2/2$, we have the following inequality:

$$m_k \cdot \frac{a_k(b_k+1/b_k)+4}{(a_k+b_k+1/b_k)^2} \leq m_k \cdot \frac{(a_k+b_k+1/b_k)^2/2+4}{(a_k+b_k+1/b_k)^2}.$$

Therefore, to show Eq. (78), we only need to show

$$m_k\left(\frac{1}{2}+\frac{4}{(a_k+b_k+1/b_k)^2}\right)\leq 2-\gamma+\iota_k.$$

Since $b_k+1/b_k\geq 2$ and $a_k\geq\exp\left(\frac{\|\mu\|_2^2}{8}\right)$, it suffices to ensure

$$m_k\leq\frac{(4-2\gamma+2\iota_k)\left(2+\exp\left(\frac{\|\mu\|_2^2}{8}\right)\right)^2}{8+\left(2+\exp\left(\frac{\|\mu\|_2^2}{8}\right)\right)^2}.$$

On the other hand, in order to show $h_\gamma(v_k,k)\geq 0$ for all $v_k\geq 0$, we only need to prove

$$m_k\cdot\frac{a_k(b_k+1/b_k)+4}{(a_k+b_k+1/b_k)^2}\geq\gamma+\iota_k. \tag{79}$$

Note that $a_k(b_k+1/b_k)\geq 2\exp(\|\mu\|^2/8)$ and $(a_k+b_k+1/b_k)^2\leq\{\exp(\|\mu\|^2/2)+2\exp(|v_k|)\}^2$, then to establish Eq. (79), it suffices to have

$$m_k\geq\frac{(\gamma+\iota_k)(\exp(\|\mu\|^2/2)+2\exp(|v_k|))^2}{2\exp(\frac{\|\mu\|^2}{8})+4}.$$

The existence of γ is ensured by making ι_k sufficiently small. For instance, we can set

$$\iota_k=\frac{(4\exp(\frac{\|\mu\|^2}{8})+8)\left(2+\exp\left(\frac{\|\mu\|_2^2}{8}\right)\right)^2}{\left[8+\left(2+\exp\left(\frac{\|\mu\|_2^2}{8}\right)\right)^2\right](\exp(\|\mu\|^2/2)+2\exp(|v_k|))^2},$$

which implies

$$\frac{\iota_k(\exp(\|\mu\|^2/2)+2\exp(|v_k|))^2}{2\exp(\frac{\|\mu\|^2}{8})+4}=\frac{2\left(2+\exp\left(\frac{\|\mu\|_2^2}{8}\right)\right)^2}{8+\left(2+\exp\left(\frac{\|\mu\|_2^2}{8}\right)\right)^2}<\frac{(4+2\iota_k)\left(2+\exp\left(\frac{\|\mu\|_2^2}{8}\right)\right)^2}{8+\left(2+\exp\left(\frac{\|\mu\|_2^2}{8}\right)\right)^2}.$$

The proof of the first claim is complete, by plugging in $\iota_k=e^{\delta_k}-1$.

Proof the second claim We denote $s(v_k,k)=\exp(e^{-T+t_k}v_k)+\exp(-e^{-T+t_k}v_k)$, which is naturally lower bounded by 2. Revisiting Eq. (76), we have

$$\begin{aligned}\frac{\partial h}{\partial v_k}(v_k,k)&=\iota_k\eta e^{-2T+2t_k}\|\mu\|_2^2\frac{\exp(e^{-2T+2t_k}\|\mu\|_2^2/2)s(v_k,k)+4}{[\exp(e^{-2T+2t_k}\|\mu\|_2^2/2)+s(v_k,k)]^2}-2-\iota_k\\&\geq\frac{1}{4}\iota_k\eta\|\mu\|_2^2\frac{\exp(\|\mu\|_2^2/8)s(v_k,k)}{\exp(\|\mu\|_2^2)+2s(v_k,k)\exp(\|\mu\|_2^2/2)+s(v_k,k)^2}-2-\max_k\iota_k\\&\geq\frac{1}{4}(\min_{0\leq k\leq K-1}\iota_k)\eta\|\mu\|_2^2\frac{\exp(\|\mu\|_2^2/8)}{\frac{\exp(\|\mu\|_2^2)}{s(v_k,k)}+s(v_k,k)+2\exp(\|\mu\|_2^2/2)}-2-\max_k\iota_k.\end{aligned}\tag{80}$$

When $\iota_k\neq 0$ for all $k\in\{0\}\cup[K-1]$, the lower bound in the display above first increases then decreases as $s(v_k,k)$ increases from 0 to ∞. We take η_0' sufficiently large, such that for any $\eta>\eta_0'$, there exists $s_0\geq 2$ dependent on $(\eta,\|\mu\|,\{\iota_k\}_{k\in\{0\}\cup[K-1]})$, such that $\frac{\partial h}{\partial v_k}(v_k,k)>0$ for all $2\leq s(v_k,k)\leq s_0$ and all $k\in\{0,1,\cdots,K-1\}$. In fact, we can choose η_0' large enough so that $\frac{\partial h}{\partial v_k}(v_k,k)|_{v_k=2}>0$. In this case, we may set s_0 to be the larger solution to the following quadratic equation (with the variable being s):

$$\frac{1}{4}(\min_{0\leq k\leq K-1}\iota_k)\eta\|\mu\|_2^2\frac{\exp(\|\mu\|_2^2/8)}{\frac{\exp(\|\mu\|_2^2)}{s}+s+2\exp(\|\mu\|_2^2/2)}-2-\max_k\iota_k=0. \tag{81}$$

One can verify that in order to have $\frac{\partial h}{\partial v_k}(v_k, k)|_{v_k=2} > 0$, it suffices to choose

$$\eta_0' \geq \frac{16 + 16\exp(\|\mu\|_2^2/2) + 8\exp(\|\mu\|_2^2) + 8\max_k \iota_k}{\|\mu\|_2^2 \exp(\|\mu\|_2^2/8) \min_{0 \leq k \leq K-1} \iota_k}. \tag{82}$$

The larger solution to Eq. (81) takes the form:

$$s_0 = \frac{1}{8(2 + \max_k \iota_k)}(\min_{0 \leq k \leq K-1} \iota_k)\eta\|\mu\|_2^2 \exp(\|\mu\|_2^2/8) - \exp(\|\mu\|_2^2/2)$$
$$+ \frac{\sqrt{\left[\frac{1}{4(2+\max_k \iota_k)}(\min_{0 \leq k \leq K-1} \iota_k)\eta\|\mu\|_2^2 \exp(\|\mu\|_2^2/8) - 2\exp(\|\mu\|_2^2/2)\right]^2 - 4\exp(\|\mu\|_2^2)}}{2}.$$

To ensure $s_0 \geq 2$, we may choose η_0' satisfying

$$\frac{1}{8(2 + \max_k \iota_k)}(\min_{0 \leq k \leq K-1} \iota_k)\eta_0'\|\mu\|_2^2 \exp(\|\mu\|_2^2/8) - \exp(\|\mu\|_2^2/2) \geq 2,$$

which gives rise to

$$\eta_0' \geq \frac{(2 + \max_k \iota_k)(16 + 8\exp(\|\mu\|_2^2/2))}{\|\mu\|_2^2 \exp(\|\mu\|_2^2/8) \min_{0 \leq k \leq K-1} \iota_k}. \tag{83}$$

Combining Eq. (82) with Eq. (83) leads to

$$\eta_0' \geq \frac{16(2 + \max_k \iota_k)(1 + \exp(\|\mu\|_2^2/2)) + 8\exp(\|\mu\|_2^2) + 8\max_k \iota_k}{\|\mu\|_2^2 \exp(\|\mu\|_2^2/8) \min_{0 \leq k \leq K-1} \iota_k}.$$

We observe that $\eta_0' \geq \eta_0$ (recall $\eta_0 = \frac{1}{\|\mu\|_2^2 \max_k \iota_k}$), and also s_0 increases linearly as the guidance strength η increases.

Similar to the derivation of Eq. (80), we get

$$\frac{\partial h}{\partial v_k}(v_k, k) \leq (\max_{0 \leq k \leq K-1} \iota_k)\eta\|\mu\|_2^2 \frac{\exp(\|\mu\|_2^2/2)s(v_k, k) + 4}{\exp(\|\mu\|_2^2/4) + 2s(v_k, k)\exp(\|\mu\|_2^2/8) + s(v_k, k)^2} - 2 - \min_k \iota_k.$$

For a sufficiently large s_1, it holds that $\frac{\partial h}{\partial v_k}(v_k, k) < 0$ whenever $s(v_k, k) > s_1$. Indeed, we can solve for s_1 explicitly as

$$s_1 = \frac{1}{2(2 + \min_k \iota_k)}(\max_k \iota_k)\eta\|\mu\|_2^2 \exp(\|\mu\|_2^2/2) - \exp(\|\mu\|_2^2/8)$$
$$+ \frac{\sqrt{\left[\frac{\max_k \iota_k}{2+\min_k \iota_k}\eta\|\mu\|_2^2 \exp(\|\mu\|_2^2/2) - 2\exp(\|\mu\|_2^2/8)\right]^2 - 4\exp(\|\mu\|_2^2/4) + 16\frac{\max_k \iota_k}{2+\min_k \iota_k}\eta\|\mu\|_2^2}}{2}.$$

Again s_1 increases as η increases, and we can ensure $s_1 \geq 2$ by choosing sufficiently large η_0'. In fact, we only require

$$\eta_0' \geq \frac{4(2 + \min_k \iota_k) + 2(2 + \min_k \iota_k)\exp(\|\mu\|_2^2/8)}{\|\mu\|_2^2 \exp(\|\mu\|_2^2/2) \max_k \iota_k}.$$

Given s_0 and s_1, we solve for a constant a so that $s(v_k, k) \leq s_0$ when $v_k \leq a$ for all k. This is plausible since by assumption e^{-T+t_k} takes value inside the interval $[1/2, 1]$. We also solve for b' so that $s(v_k, k) \geq s_0$ when $v_k \geq b'$ for all k. Checking the definition of $s(v_k, k)$, we conclude that we can choose a, b' appropriately, such that both of them increase as s_0 and s_1 increase, respectively. Recall that both s_0 and s_1 are increasing functions of η. Therefore, we deduce that we can find a and b' that satisfy all the above desiderata. Furthermore, both of them get larger as we increase the guidance strength η.

To summarize, we conclude that for any $\eta \geq \eta_0'$, $h(v_k, k)$ is strictly increasing for all k when $v_k \in [0, a]$, and strictly decreasing for all k when $v_k \geq b'$. Since $h(v_k, k)$ is continuous in v_k and $h(\infty, k) < 0$, there exists $b > 0$ such that $h(v_k, k) < 0$ when $v_k \geq b$ for all k. Hence, we have established that for all possible k, it holds that $h(v_k, k) > 0$ for

$v_k \in (0, a]$ and $h(v_k, k) < 0$ for $v_k > b$. An illustration of the $h(v_k, k)$ curve can be found in the right panel of Figure 4. Next, we apply the same argument that we used to derive the first claim, and deduce that

$$\begin{aligned} |v_{k+1}| > |v_k| \quad &\text{if} \quad |v_k| \le a; \\ |v_{k+1}| < |v_k| \quad &\text{if} \quad |v_k| \ge b. \end{aligned}$$

We skip the proof details for the equations above to avoid redundancy. The second claim is verified and thus the proof is complete. □

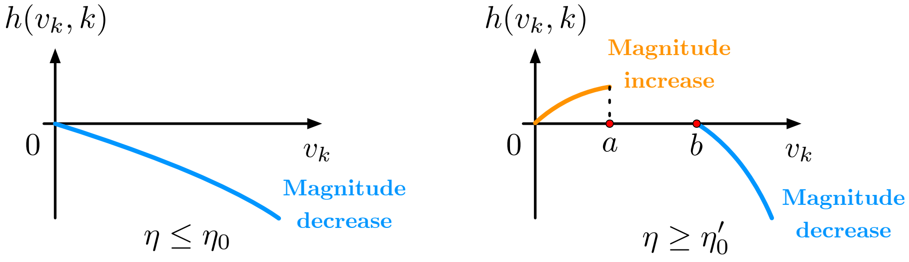

Figure 4. Illustration of the behaviors of $h(v_k, k)$ when constrained to the positive real line, under different ranges of guidance strength η. The left panel corresponds to a small strength $\eta < \eta_0$. In this case, $h(v_k, k)$ is negative and decreasing for all $v_k \ge 0$. In contrast, the right panel corresponds to a large strength $\eta > \eta_0'$, where $h(v_k, k)$ is increasing on $[0, a]$ and decreasing on $[b, \infty)$.

D Additional numerical experiments

We collect in this section outcomes from additional numerical experiments. We first consider discretized samplers, and verify the theoretical results in Section 5.2. Specifically, Figures 5 and 6 demonstrate the behavior of DDIM in 2D/ 3D symmetric 3-component GMMs, Figures 7 and 8 display the corresponding behavior of DDPM. As our theory (Proposition 5.3) suggests, when the guidance strength is enormously large, the middle component splits into two clusters. Such phenomenon is not limited to symmetric GMMs: as shown by Figures 9 and 10, for a large enough guidance strength, the middle component becomes distorted under diffusion guidance even in the context of a non-symmetric GMM.

We then switch to continuous-time samplers, and the goal is in turn to justify the theoretical implications listed in Section 3. More precisely, in Figure 11 we visualize the effect of diffusion guidance on a 3-component symmetric GMM in $\mathbb{R}^3$. This can be regarded as an analogue of Figure 1 in the 3D setting. We further confirm our theoretical results by Figure 12, which demonstrates how classification confidence and differential entropy evolve as guidance strength η increases.

Finally, we confirm our theoretical findings using real-world datasets.

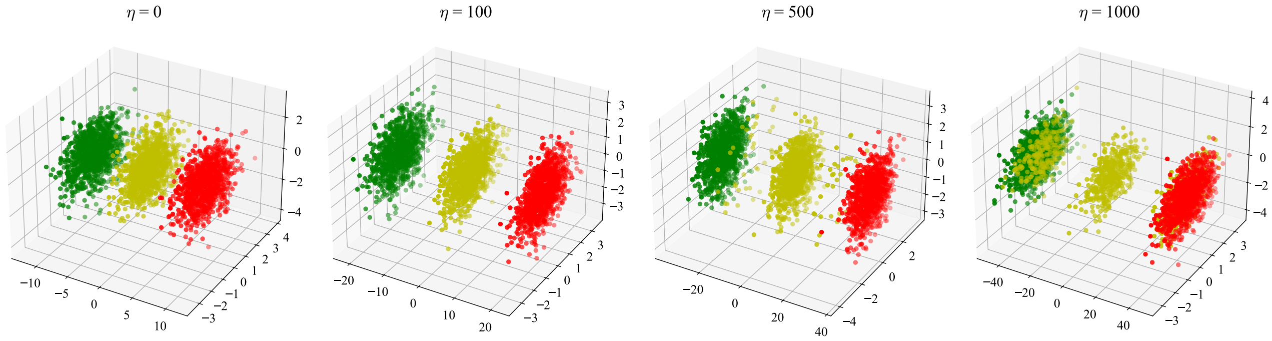

Figure 5. Illustration of the effect of guidance on a discretized DDIM sampler. For this experiment, we set $p_* = \frac{1}{3}\mathsf{N}((0,0,0), I_3) + \frac{1}{3}\mathsf{N}\big((0,\sqrt{3},0), I_3\big) + \frac{1}{3}\mathsf{N}\big((0,-\sqrt{3},0), I_3\big)$, $T = 10$, and $\delta_k = 0.01$ for all possible k. From the plot, we see that the middle component splits with a sufficiently large η. To summarize, the numerical observations corroborate our theory.

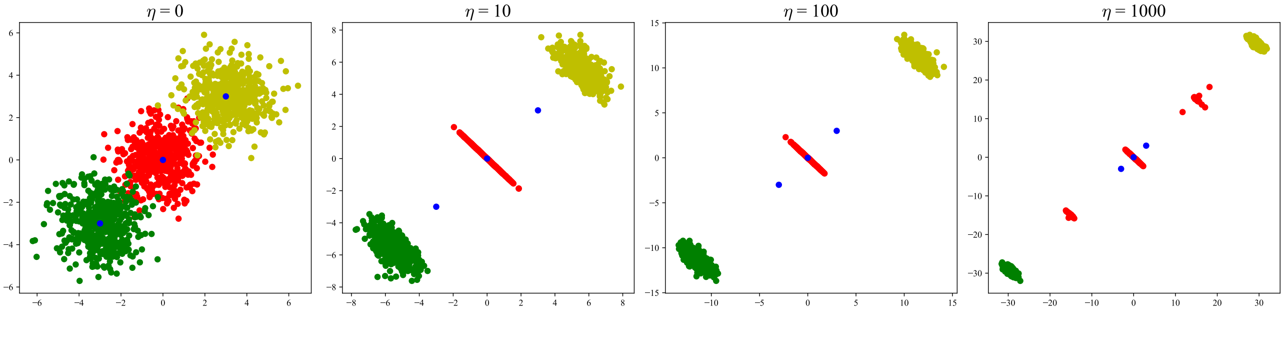

Figure 6. Illustration of the effect of guidance on a discretized DDIM sampler. For this experiment, we set $p_* = \frac{1}{3}\mathsf{N}((0,0), I_2) + \frac{1}{3}\mathsf{N}\big((3,3), I_2\big) + \frac{1}{3}\mathsf{N}\big((-3,-3), I_2\big)$, $T = 10$, and $\delta_k = 0.01$ for all possible k. The same splitting phenomenon is observed with a sufficiently large η.

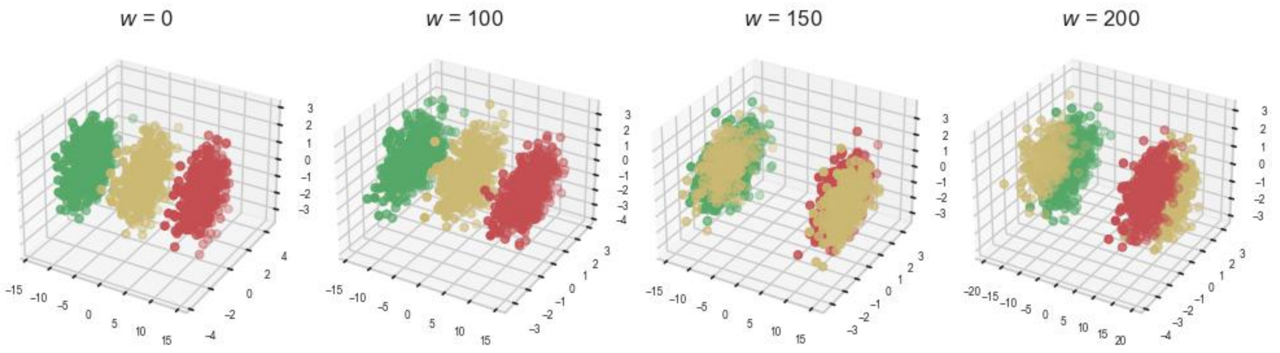

Figure 7. Illustration of the effect of guidance on a discretized DDPM sampler. The experiment setup is the same as that of Figure 5.

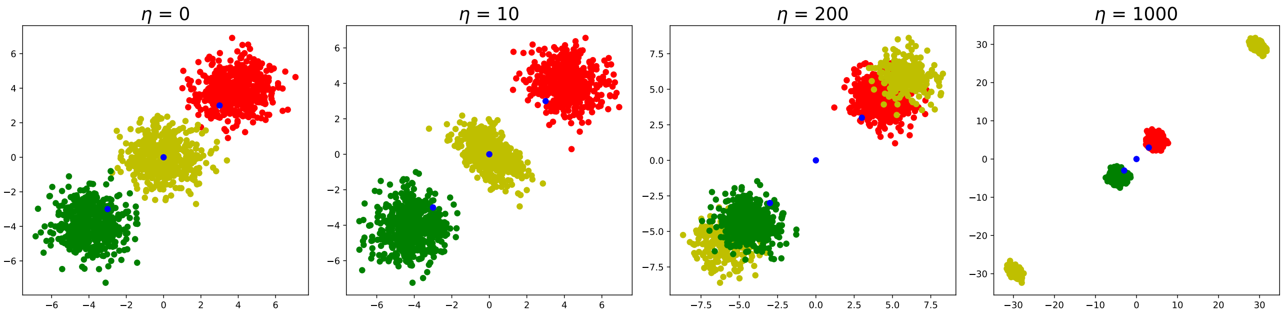

Figure 8. Illustration of the effect of guidance on a discretized DDPM sampler. The experiment setup is the same as that of Figure 6.

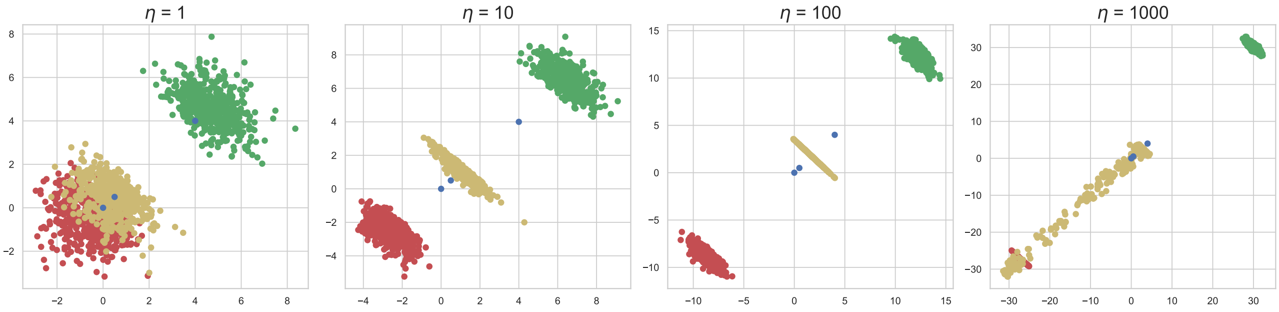

Figure 9. Illustration of the effect of guidance on a discretized DDPM sampler. Here, $p_* = \frac{1}{3}\mathsf{N}((0,0), I_2) + \frac{1}{3}\mathsf{N}\big((0.5, 0, 5), I_2\big) + \frac{1}{3}\mathsf{N}\big((4,4), I_2\big)$, $T = 10$, and $\delta_k = 0.01$ for all possible k. In this asymmetric GMM, the center component penetrates the left side component (the left component is colored in red, and without any guidance is close to the center component) under large enough guidance.

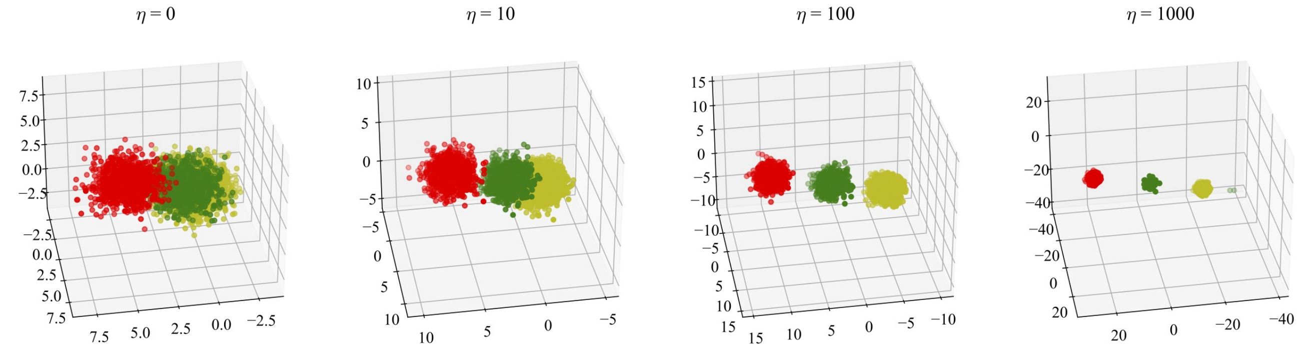

Figure 10. Illustration of the effect of guidance on a discretized DDIM sampler. Here, $p_* = \frac{1}{3}\mathsf{N}((0,0,0), I_3) + \frac{1}{3}\mathsf{N}\big((0.5, 0.5, 0), I_3\big) + \frac{1}{3}\mathsf{N}\big((5,5,0), I_3\big)$, $T = 10$, and $\delta_k = 0.01$ for all possible k. The observation is similar to Figure 8 for the 2D case. The center component splits into two components under sufficiently large guidance.

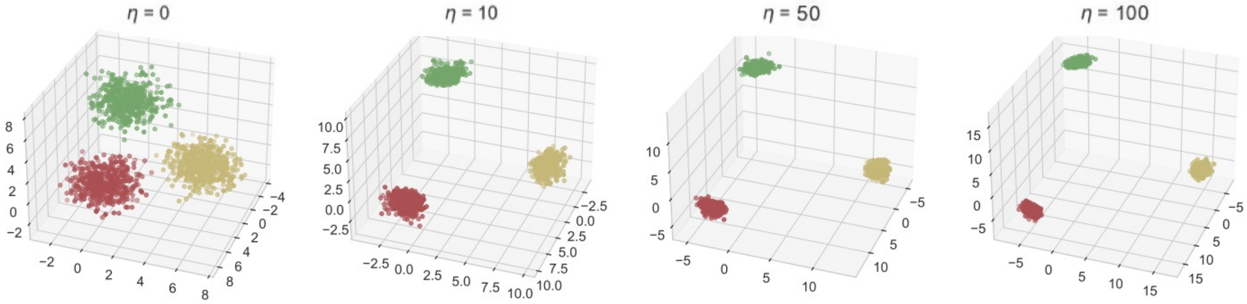

Figure 11. Illustration of the effect of guidance on a continuous-time DDIM sampler. Here, $p_* = \frac{1}{3}\mathsf{N}((1,0,0), I_3) + \frac{1}{3}\mathsf{N}\big((0,1,0), I_3\big) + \frac{1}{3}\mathsf{N}\big((0,0,1), I_3\big)$. This setting satisfies Assumption 3.1, and we observe that the components become more separated as we increase the guidance strength.

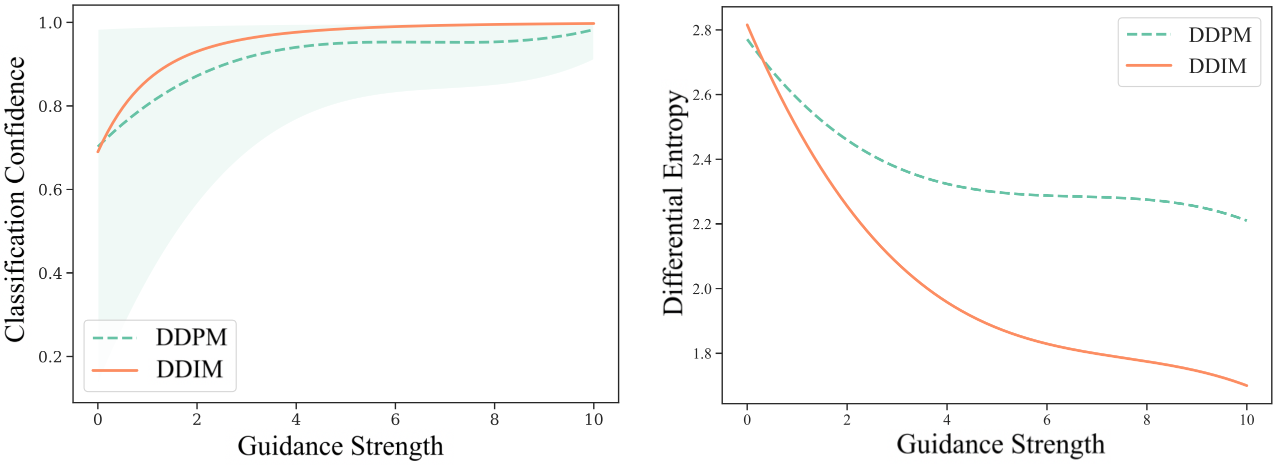

Figure 12. The effect of diffusion guidance on continuous-time samplers. We consider a three-component equidistant GMM: $p_* = \frac{1}{3}\mathsf{N}((0,1), I_2) + \frac{1}{3}\mathsf{N}\big((\frac{\sqrt{3}}{2}, -\frac{1}{2}), I_2\big) + \frac{1}{3}\mathsf{N}\big((-\frac{\sqrt{3}}{2}, -\frac{1}{2}), I_2\big)$. (a) In the left panel, we initiate the reverse processes at the origin and record the classification confidence (measured by the posterior probability of class label) under different levels of guidance. For the SDE-based sampler, the output sample is random. We generate 10^3 samples for each guidance strength and plot the 97.5% and 2.5% quantiles. (b) In the right panel, we initiate the processes following a standard Gaussian distribution and plot the differential entropy of the output distributions. For each guidance strength, we generate 10^4 samples. We adopt the function `scipy.neighbors.KernelDensity` from the `scipy` module in Python to estimate the density function of the generated distribution using one half of the generated samples, and use the other half for a Monte Carlo algorithm to estimate the differential entropy.

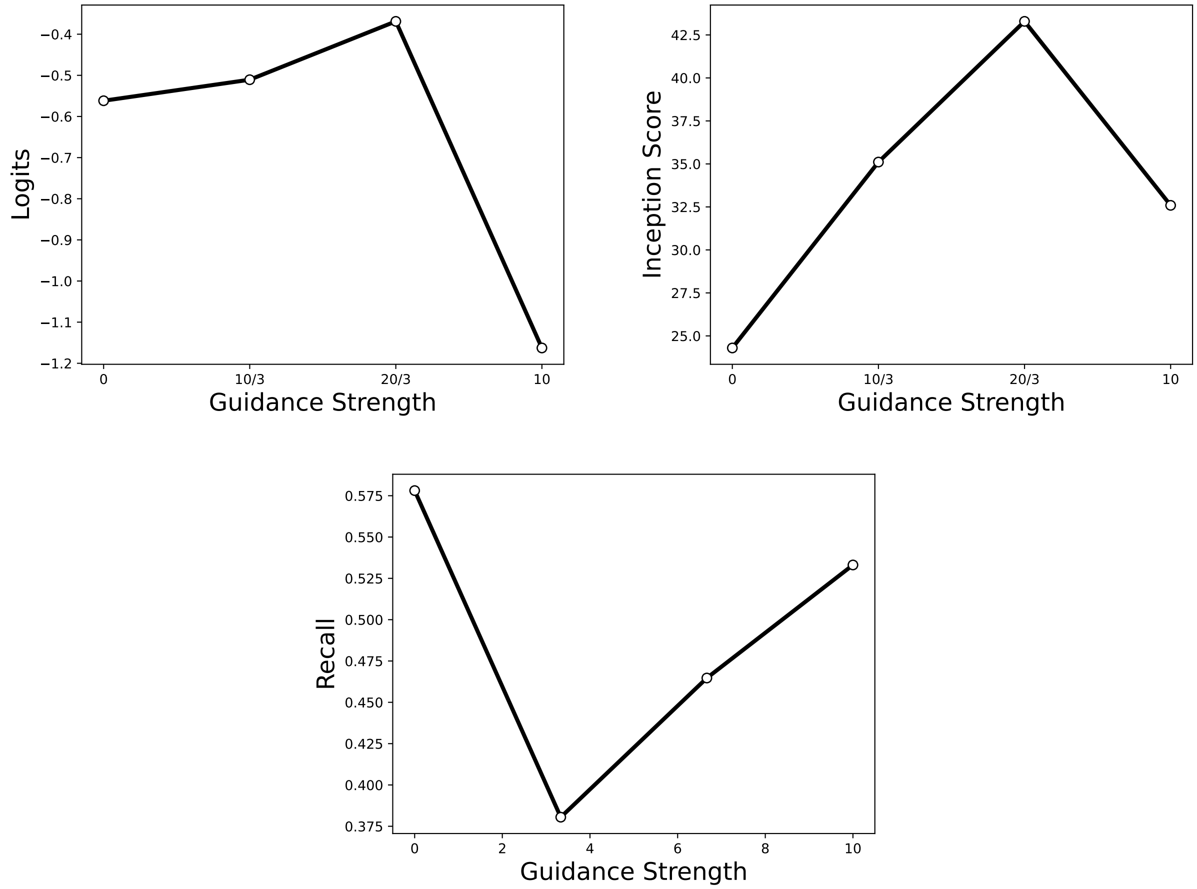

Figure 13. We illustrate the effect of guidance on classification confidence on ImageNet 64 × 64. Here, we reproduce the classifier guidance in (Dhariwal & Nichol, 2021). We vary guidance strength η in the set $\{0, 10/3, 20/3, 10\}$. We generate images by first randomly sample an image label, and then conditioned on the label, we generate an image using the diffusion model. Once an image is generated, we compute its classification confidence via a pre-trained classifier (Inception-V3). We report (1) the average logit (logarithm of classification likelihood), (2) the Inception Score (Kynkäänniemi et al., 2019), which measures sample fidelity, and (3) the Recall metric (Kynkäänniemi et al., 2019), which measures sample diversity. All these metrics are calculated over 100 randomly generated images. From the plot, we see that classification confidence and fidelity increase and the diversity decreases as we increase guidance strength from 0 to $20/3$, which corroborates our theoretical results in Theorem 5.1. Note that the classification confidence and fidelity decrease and diversity increase for $\eta = 10$, suggesting the emergence of negative effects due to overly strong guidance.