IMPROVED DDIM SAMPLING WITH MOMENT MATCHING GAUSSIAN MIXTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose using a Gaussian Mixture Model (GMM) as reverse transition operator (kernel) within the Denoising Diffusion Implicit Models (DDIM) framework, which is one of the most widely used approaches for accelerated sampling from pre-trained Denoising Diffusion Probabilistic Models (DDPM). Specifically we match the first and second order central moments of the DDPM forward marginals by constraining the parameters of the GMM. We see that moment matching is sufficient to obtain samples with equal or better quality than the original DDIM with Gaussian kernels. We provide experimental results with unconditional models trained on CelebAHQ and FFHQ and class-conditional models trained on ImageNet datasets respectively. Our results suggest that using the GMM kernel leads to significant improvements in the quality of the generated samples when the number of sampling steps is small, as measured by FID and IS metrics. For example on ImageNet 256x256, using 10 sampling steps, we achieve a FID of 6.94 and IS of 207.85 with a GMM kernel compared to 10.15 and 196.73 respectively with a Gaussian kernel.

1 INTRODUCTION

Diffusion models (Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021), also known as Score based generative models (Sohl-Dickstein et al., 2015; Song et al., 2020), have demonstrated great success in modeling data distributions in various domains including images (Dhariwal & Nichol, 2021; Nichol & Dhariwal, 2021; Saharia et al., 2022; Rombach et al., 2022), videos (Ho et al., 2022; Blattmann et al., 2023), speech (Kong et al., 2021) and 3D (Poole et al., 2022; Watson et al., 2022). This is due to their flexibility in modeling complex multimodal distributions and ease of training relative to other competitive approaches such as VAEs (Kingma & Welling, 2014; Rezende et al., 2014), GANs (Goodfellow et al., 2014; Salimans et al., 2016; Karras et al., 2018; Brock et al., 2019), autoregressive models (van den Oord et al., 2016b;a) and normalizing flows (Rezende & Mohamed, 2015; Dinh et al., 2017), which do not exhibit both of the advantages simultaneously. In spite of their success, the main bottleneck to their adoption is the slow sampling speed, usually requiring hundreds to thousands of denoising steps to generate a sample. Recently a number of accelerated sampling approaches have emerged, notably Denoising Diffusion Implicit Models (Song et al., 2021), pseudo numerical methods (Liu et al., 2022), distillation (Salimans & Ho, 2022; Luhman & Luhman, 2021; Meng et al., 2023) and various approximations to solving the reverse SDE(Song et al., 2020; Lu et al., 2022; 2023; Zhang & Chen, 2023).

Denoising Diffusion Implicit Models (DDIM) (Song et al., 2021) accelerate sampling from Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020) by hypothesizing a family of non-Markovian forward processes, whose reverse process (Markovian) estimators can be trained with the same surrogate objective as DDPMs, assuming the same parameterization for reverse estimators. It is based on the observation that a simplified DDPM training objective $\mathcal{L}_{simple,w}$ (Eq. 4) depends on the forward process only through its marginals at discrete steps t. In other words, one can sample with a pretrained DDPM denoiser by designing a different forward/backward process than the original DDPM given that the forward marginals are the same. Recent works (Xiao et al., 2022; Guo et al., 2023) have shown that the true denoising conditional distributions of DDPMs are multimodal, especially when the denoising step sizes are large. Although the unimodal Gaussian kernel in DDIM yields a multimodal denoising conditional distribution, there is potential to improve its expressiveness with a multimodal kernel. It is not straightforward to use a multimodal kernel

while satisfying the marginal constraints. A recent work (Watson et al., 2021) shows that it is not necessary to satisfy the marginal constraints to achieve accelerated sampling. Taking inspiration from these works, we propose using Gaussian mixtures as the transition kernels of the reverse process in the DDIM framework. This results in a non-Markovian inference process with Gaussian mixtures as marginals. Further, we constrain the mixture parameters so that the first and second order central moments of the forward marginals match exactly those of the DDPM forward marginals. For brevity, we refer to *central moments* as simply *moments* in the rest of the paper. In summary, our main contributions are as follows:

1. We propose using Gaussian Mixture Models (GMM) as reverse transition operators (kernel) within the DDIM framework, which results in a non-Markovian inference process with Gaussian mixtures as marginals.
2. We derive constraints to match the first and second order moments of the resulting forward GMM marginals to those of the DDPM forward marginals. Based on these constraints, we provide three different schemes to compute GMM parameters efficiently.
3. We demonstrate experimentally that the proposed method results in further accelerating sampling from pretrained DDPM models relative to DDIM, especially with fewer sampling steps.

We begin by providing a brief overview of Diffusion and the DDIM framework in Section 2. Our approach for extending DDIMs with Gaussian mixture transition kernels is provided in Section 3. In Section 4, we conduct experiments on CelebAHQ (Liu et al., 2015), FFHQ (Karras et al., 2019), ImageNet (Deng et al., 2009), and text-to-image generation with Stable Diffusion (Rombach et al., 2022), and provide quantitative results. Finally we conclude in Section 5.

2 Background

2.1 Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models (Ho et al., 2020) learn a model for the data distribution $q(\boldsymbol{x}_0)$ by designing a forward and backward diffusion process. In the forward process, noise is added to the data samples following a predetermined schedule, thereby transforming a structured distribution $q(\boldsymbol{x}_0)$ at step $t = 0$ to Gaussian noise, $q(\boldsymbol{x}_T) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, at step T. This is achieved by setting up a Markov chain at every step t with the transition kernel defined by

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}\left(\sqrt{\frac{\alpha_t}{\alpha_{t-1}}}\boldsymbol{x}_{t-1}, \left(1 - \frac{\alpha_t}{\alpha_{t-1}}\right)\boldsymbol{I}\right), \tag{1}$$

where α_t are chosen such that the marginal $q(\boldsymbol{x}_T)$ converges to $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ with large enough T. It is straightforward to obtain the marginal of the latent $\boldsymbol{x}_t$ at any step t conditioned on data sample $\boldsymbol{x}_0$ as

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}\left(\sqrt{\alpha_t}\boldsymbol{x}_0, (1 - \alpha_t)\boldsymbol{I}\right). \tag{2}$$

In the backward process, a parameterized Markovian denoiser $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$, initialized with Gaussian noise, $p(\boldsymbol{x}_T) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, learns to estimate the distribution of $\boldsymbol{x}_{t-1}$ given $\boldsymbol{x}_t$ to maximize the evidence lower bound (ELBO) or minimize $\mathcal{L}_{ELBO}$:

$$\mathcal{L}_{ELBO} = \mathbb{E}_q\left[D_{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)||p_\theta(\boldsymbol{x}_T)) + \sum_{t=2}^{T} D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)||p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) - \log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1))\right]. \tag{3}$$

The above is equivalent to training $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ to match the posterior $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$, which turns out to be a Gaussian (Luo, 2022). Assuming a Gaussian form for $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}|\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t))$, leads to a simplified loss function for training DDPMs, which is optimized over $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ and $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$. It has been found that a reparameterized mean estimator $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ as a function of added noise estimator $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ has benefits of better sample quality. The resulting simplified loss function $\mathcal{L}_{simple,w}$ (Ho et al., 2020) is a weighted version of ELBO with weights w_t:

$$\mathcal{L}_{simple,w} = \mathbb{E}_{t\sim U[1,T]q(\boldsymbol{x}_0)q(\boldsymbol{x}_t|\boldsymbol{x}_0)\boldsymbol{\epsilon}}\left[w_t||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)||^2\right]. \tag{4}$$

$\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ is modeled as a U-Net (Ho et al., 2020) or a transformer (Peebles & Xie, 2022). The covariance estimator $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$ is either learned (Nichol & Dhariwal, 2021; Dhariwal & Nichol, 2021) or fixed (Ho et al., 2020).

2.2 DDIM

Following the same notation as Song et al. (2021), we refer to the forward process as *inference* and the reverse process as *generative*. The key assumption is that the more general non-Markovian inference processes $q_\sigma(\boldsymbol{x}_{0:T})$ have the same marginal distribution $q_\sigma(\boldsymbol{x}_t|\boldsymbol{x}_0)$ at every t as the DDPM, but not necessarily the same joint distribution over all the latents $q_\sigma(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)$. The form of a Markovian family of generative process in DDIMs (Song et al., 2021) is given by

$$\begin{aligned} q_\sigma(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) &\coloneqq q_\sigma(\boldsymbol{x}_T|\boldsymbol{x}_0) \prod_{t=2}^{t=T} q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0), \\ q_\sigma(\boldsymbol{x}_T|\boldsymbol{x}_0) &\coloneqq \mathcal{N}\left(\sqrt{\alpha_T}\boldsymbol{x}_0, (1-\alpha_T)\boldsymbol{I}\right), \end{aligned} \tag{5}$$

where $\sigma \in R^T_{\geq 0}$ parameterizes the variances of the reverse transition kernels,

$$q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}\left(\sqrt{\alpha_{t-1}}\boldsymbol{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}.\frac{\boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_0}{\sqrt{1-\alpha_t}}, \sigma_t^2\boldsymbol{I}\right), \forall t > 1. \tag{6}$$

The transition kernel above ensures that the resulting marginals $q_\sigma(\boldsymbol{x}_t|\boldsymbol{x}_0)$ are identical to the DDPM marginals in Eq. 2. The ODE perspective of DDIM and other related works on accelerated sampling from diffusion models are discussed in Appendix A.1.

3 Approach

We propose using a Gaussian Mixture Model (GMM) within the reverse transition kernels of the DDIM generative process. Specifically, the form of transition kernels in Eq. 6 is given by

$$q_{\sigma,\mathcal{M}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \sum_{k=1}^{K} \pi_t^k \mathcal{N}\left(\sqrt{\alpha_{t-1}}\boldsymbol{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}.\frac{\boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_0}{\sqrt{1-\alpha_t}} + \boldsymbol{\delta}_t^k, \sigma_t^2\boldsymbol{I} - \boldsymbol{\Delta}_t^k\right), \tag{7}$$

where $\mathcal{M}_t = \left(\pi_t^k, \boldsymbol{\delta}_t^k, \boldsymbol{\Delta}_t^k\right), k = 1 \dots K$ denote the additional GMM parameters, specifically the mixture component priors, mean and covariance offsets relative to the single Gaussian counterparts of Eq. 6, respectively. Further, we constrain the above kernel so that the first and second order moments of the individual latent variables $q_{\sigma,\mathcal{M}}(\boldsymbol{x}_t|\boldsymbol{x}_0)$ are the same as that of an equivalently parameterized DDPM (Eq. 2). This allows us to use DDIM sampling on a model trained with the same surrogate objective as the DDPM in Eq. 4 given that the GMM parameters $\mathcal{M}_t$ satisfy:

$$\begin{aligned} &\sum_{k=1}^{K} \pi_t^k = 1, \;\; \sum_{k=1}^{K} \pi_t^k \boldsymbol{\delta}_t^k = 0 \\ &\boldsymbol{\Delta}_t^k = \boldsymbol{\delta}_t^k (\boldsymbol{\delta}_t^k)^T \quad \text{OR} \quad \boldsymbol{\Delta}_t^k = \frac{1}{K\pi_t^k} \sum_{l=1}^{K} \pi_t^l \boldsymbol{\delta}_t^l (\boldsymbol{\delta}_t^l)^T, \end{aligned} \tag{8}$$

where either one of the two constraints on the covariance matrix offset $\boldsymbol{\Delta}_t^k$ is sufficient to yield the correct moment matching. Please see Appendix A.3 for proof. We also provide an upper bound for the ELBO loss using the proposed inference process as an augmented version of the $\mathcal{L}_{simple,w}$ loss in Appendix A.6. The proposed kernel yields a more expressive multimodal denoising conditional distribution $q_{\sigma,\mathcal{M}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ compared to DDIM as shown in Appendix A.7.

3.1 GMM Parameters

Sampling with the original DDIM kernel of Eq. 6 requires choosing an appropriate value for the variance σ_t^2, which determines the *stochasticity* (Song et al., 2021) of the DDIM inference and sampling processes. It is specified as a proportion η of the DDPM reverse transition kernel's variance at

the corresponding step t. The proposed approach requires choosing the additional GMM parameters $\mathcal{M}_t$ at every step t during sampling. In what follows, we describe three different ways to choose these parameters efficiently to satisfy the constraints in Eq. 8 while keeping additional computational requirements relatively low.

First we choose the mixture priors π_t^k to be uniform or with a suitable random initialization so that they are non-negative and sum to one. We experiment with choosing the mean offsets $\boldsymbol{\delta}_t^k$ either randomly (DDIM-GMM-RAND) or followed by orthogonalization (DDIM-GMM-ORTHO) to allow for better exploration of the latent space ($\boldsymbol{x}_t, t > 0$) as described below.

3.1.1 METHOD 1: DDIM-GMM-RAND

At every step t we sample random vectors $\boldsymbol{o}_t^k, k = 1 \dots K$ from an isotropic multivariate Gaussian with dimensionality equal to that of the latent variables $\boldsymbol{x}_t \in R^D$. These vectors are mean centered and scaled to yield the offsets $\boldsymbol{\delta}_k^t$.

$$\begin{aligned}
&\boldsymbol{O}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \boldsymbol{O}_t \in R^{D \times K}, K < D \\
&\bar{\boldsymbol{o}}_t = \sum_{k=1}^{K} \pi_t^k \boldsymbol{O}_t[k], \;\; \boldsymbol{C}_t[k] = \boldsymbol{O}_t[k] - \bar{\boldsymbol{o}}_t, \;\; \boldsymbol{\delta}_t^k = \frac{s}{||\boldsymbol{C}_t[k]||_2} \boldsymbol{C}_t[k],
\end{aligned} \tag{9}$$

where $\boldsymbol{O}_t[k]$ denotes the kth column of the matrix $\boldsymbol{O}_t$, $||\boldsymbol{C}_t[k]||_2$ denotes the magnitude of $\boldsymbol{C}_t[k]$, and s is a scale factor that controls the magnitude of the offsets.

3.1.2 METHOD 2: DDIM-GMM-ORTHO

In order to allow for better exploration of the latent space of $\boldsymbol{x}_t$, the set of offsets above is orthonormalized using an SVD on the matrix $\boldsymbol{O}_t$ with $\boldsymbol{o}_t^k$ as columns and choosing the first K components of the output $\boldsymbol{U}_t$ factor, i.e. the first K eigenvectors of $\boldsymbol{O}_t \boldsymbol{O}_t^T$. Specifically

$$\begin{aligned}
&\boldsymbol{U}_t \boldsymbol{\Sigma}_t \boldsymbol{V}_t^T = SVD(\boldsymbol{O}_t) \\
&\bar{\boldsymbol{u}}_t = \sum_{k=1}^{K} \pi_t^k \boldsymbol{U}_t[k], \;\; \boldsymbol{C}_t[1:K] = \boldsymbol{U}_t[1:K] - \bar{\boldsymbol{u}}_t, \;\; \boldsymbol{\delta}_t^k = s\boldsymbol{C}_t[k],
\end{aligned} \tag{10}$$

where $\boldsymbol{U}_t[1:K]$ are the first K columns of $\boldsymbol{U}_t$. The mean centering above ensures that the offsets $\boldsymbol{\delta}_t^k$ satisfy the zero weighted mean constraint in Eq. 8. The covariance parameters are chosen as

$$\boldsymbol{\Delta}_t^k = \frac{1}{K\pi_t^k} \sum_{l=1}^{K} \pi_t^l \boldsymbol{\delta}_t^l (\boldsymbol{\delta}_t^l)^T \tag{11}$$

to satisfy the covariance constraint of Eq. 8. To sample from the reverse kernel of Eq. 7, we approximate the covariance matrix $\sigma_t^2 \boldsymbol{I} - \boldsymbol{\Delta}_t^k$ to be diagonal. A straightforward approximation is to choose only the diagonal elements of $\boldsymbol{\Delta}_t^k$ and subtract from σ_t^2.

3.1.3 METHOD 3: DDIM-GMM-ORTHO-VUB

Given the random choice of offsets $\boldsymbol{\delta}_t^k$, we also experiment with an upper bound diagonal approximation of $\boldsymbol{\Delta}_t^k$ by eigen decomposition similar to PCA (Jolliffe, 1986), These variance upper bounds (VUB) determine the maximum allowable variances for the dimensions in $\boldsymbol{\Delta}_t^k$ keeping the total variance the same. If $\lambda_i, i = 1 \dots K$ are the eigenvalues of $\boldsymbol{\Delta}_t^k$, then

$$\begin{aligned}
\frac{s^2}{K\pi_t^k} \left(\pi_t^1 - \sum_{l=1}^{K} (\pi_t^l)^2 \right) \leq \lambda_1 \leq \frac{s^2}{K\pi_t^k} \pi_t^1 \\
\frac{s^2}{K\pi_t^k} \pi_t^{i-1} \leq \lambda_i \leq \frac{s^2}{K\pi_t^k} \pi_t^i, \qquad i = 2 \dots K.
\end{aligned} \tag{12}$$

Please see Appendix A.4 for proof. It turns out the upper bounds above are independent of the $\boldsymbol{\delta}_t^k$'s. We use the upper bounds in Eq. 24 to compute the diagonal approximation of $\sigma_t^2 \boldsymbol{I} - \boldsymbol{\Delta}_t^k$ by offsetting the first K elements of $\sigma_t^2 \boldsymbol{I}$ with the eigenvalue upper bounds. The scale s can be

chosen such that the upper bounds are always smaller than σ_t^2 to ensure positive variances across all dimensions. Note that the above diagonalization and variance offsetting has to be done only once before sampling, which introduces additional computation for initialization but not during sampling. We also experiment with sharing GMM parameters across sampling steps t to save time by avoiding the expensive SVD operation for each step and doing it only once. We denote this approach as DDIM-GMM-ORTHO-VUB*. Please see Appendix A.10 for further discussion on additional computational overhead.

4 EXPERIMENTS

In this section we compare the quality of samples generated using the proposed approach with those generated by the original DDIM sampling. We conduct experiments on CelebAHQ (Liu et al., 2015) and FFHQ (Karras et al., 2019), which are high resolution face datasets used as standard benchmarks for evaluating generative models. We also evaluate the effectiveness of the proposed approach on sampling from class-conditional distributions by training on the ImageNet dataset with conditioning on class labels (Rombach et al., 2022). The sample quality is measured using Frechét Inception Distance (FID) (Heusel et al., 2017) and Inception score (IS) (Salimans et al., 2016) for class-conditional generation. We train diffusion models using the unweighted DDPM objective (Ho et al., 2020) in the latent space of a VQVAE (Rombach et al., 2022). More experimental details can be found in the Appendix A.2. For each dataset, we generate as many samples as in the standard validation split of the dataset to compute the FID and IS metrics, i.e., 5000 for CelebAHQ, 10000 for FFHQ and 50000 for ImageNet respectively.

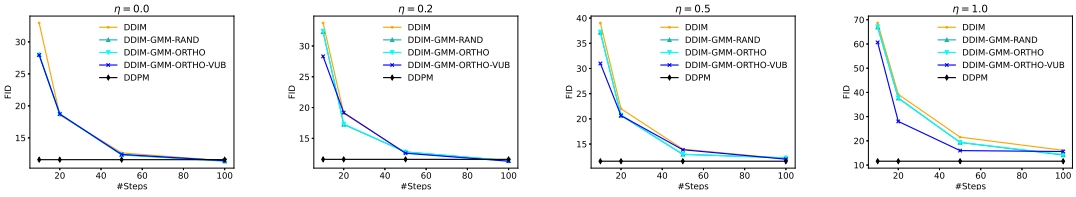

Figure 1: **CelebAHQ**. FID (↓). The horizontal line is the DDPM baseline run for 1000 steps.

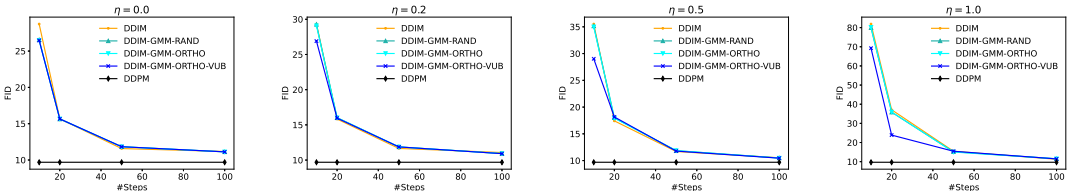

Figure 2: **FFHQ**. FID (↓). The horizontal line is the DDPM baseline run for 1000 steps.

4.1 UNCONDITIONAL MODELS ON CELEBAHQ AND FFHQ

The FID scores of unconditional generation models on CelebAHQ and FFHQ datasets are reported in Figs. 1 and 2 respectively. We run both DDIM and the proposed variants of DDIM-GMM samplers for different numbers of steps (10, 20, 50, 100) using different values of the stochasticity parameter η (Song et al., 2021). For each DDIM-GMM variant, we choose a GMM with 8 mixture components with uniform priors (π_t^k=0.125) for all steps t. We also search for the best value of scaling s among $\{0.01, 0.1, 1.0, 10.0\}$ and report the best result with the chosen value for s. For all the experiments here, we set the value of s to be the same for all steps t. It is possible to further tune these parameters. For instance one could search for an optimal set of parameters using a suitable objective (Watson et al., 2021; Mathiasen & Hvilshøj, 2021) on the training set. We also run full DDPM sampling for 1000 steps as a baseline since all the models were trained for 1000 steps in the forward process using the $\mathcal{L}_{simple,w}$ DDPM objective (Eq. 4) with uniform weights ($w = \mathbf{1}$).

We observe that sampling with a GMM transition kernel (DDIM-GMM-*) shows significant improvements in sample quality over the Gaussian kernel (DDIM) at lower values of sampling steps and higher values of η for unconditional generation on both CelebAHQ and FFHQ (see also Tables 5 and 6, Appendix A.13). Among the different choices for computing GMM offset parameters, DDIM-GMM-RAND and DDIM-GMM-ORTHO produce similar quality results. We observe significant improvements with upper bounding variances with the DDIM-GMM-ORTHO-VUB variant. Our hypothesis is that the GMM kernel allows exploring the latent space better than the Gaussian kernel under those settings. Variance upper bounding further encourages this by lumping variances into fewer dimensions of Δ_t^k. This is favorable since sampling time is a significant bottleneck for the use of DDPM in real-time applications.

4.2 Class-conditional ImageNet

We train class-conditional models on ImageNet and experiment with guided sampling using either classifier guidance (Dhariwal & Nichol, 2021) or classifier-free guidance (Ho & Salimans, 2021) below. See Appendix A.8 for additional results without using any guidance.

4.2.1 Class-conditional ImageNet with Classifier Guidance

For classifier guidance, we train a separate classifier at different levels of noise and use it with two guidance scales (1, 10) for sampling with 10 and 100 steps. The FID and IS results are in Figs. 3 and 4 respectively. Using smaller guiance scale (1), DDIM-GMM-* samplers show improvements over DDIM only under the highest $\eta(=1)$ setting. FID improves using the fewest sampling steps (10) and IS improves using both 10 and 100 sampling steps. This can be attributed to a similar argument as for unconditional sampling, especially for the least number of sampling steps. With a higher guidance scale (10), all variants of DDIM-GMM-* samplers yield significantly lower FIDs than the DDIM sampler when the number of sampling steps is small (10) (see Table 9, Appendix A.13.3). The FIDs with variance upper bounding, relative to without, improve significantly with higher values of η possibly due to greater exploration of the latent space under those settings. The differences between DDIM and DDIM-GMM-* are marginal using 100 sampling steps with the exception of DDIM-GMM-RAND and DDIM-GMM-ORTHO for the highest η setting. With a higher guidance scale (10), the IS scores of samples from DDIM-GMM-* samplers are almost always higher than from the DDIM sampler (see Table 10, Appendix A.13.3). The only exception is the DDIM-GMM-ORTHO sampler run for 100 steps using $\eta = 1$, which is only marginally worse. Similar to FID results, the differences between samplers with and without variance upper bounding are amplified by η. This is an interesting result indicating that better exploration of latent spaces with a multimodal reverse kernel not only helps with coverage (FID) but also sample sharpness (IS) since the guidance scale is known to trade-off one versus the other (Dhariwal & Nichol, 2021) and poses challenges for higher order ODE solvers (Lu et al., 2023).

4.2.2 Class-conditional ImageNet with Classifier-free Guidance

For classifier-free guidance, we jointly train a class-conditional and unconditional model with parameter sharing (Ho & Salimans, 2021) by setting the unconditional training probability to 0.1. We then sample from this model using two guidance scales (2.5, 5) for 10 and 100 sampling steps. Note that each sampling step involves two Neural Function Evaluations (NFE) using classifier-free guidance in order to compute conditional and unconditional scores with the same denoising network. The FID and IS results are shown in Figs. 5 and 6 respectively. Using fewer sampling steps (10), the FID and IS scores of the samples improve significantly when any DDIM-GMM-* sampler is used, relative to DDIM, regardless of the guidance scale (see Tables 11 and 12, Appendix A.13.3). Notably, the best FID (6.72) and IS (320.66) metrics are obtained using deterministic ($\eta = 0$) DDIM-GMM-ORTHO-VUB* and DDIM-GMM-ORTHO samplers with guidance scale 2.5 and 5.0 respectively. Similar to previous results, among DDIM-GMM-*, using variance bounding leads to more significant improvements at higher η relative to not, especially at lower guidance scale (2.5) for FID but both scales for IS.

Using 100 steps, samples from DDIM and DDIM-GMM-* variants have similar metrics in most settings with some exceptions. DDIM-GMM-RAND and DDDIM-GMM-ORTHO yield the best FID values under the $\eta = 1$ setting at both guidance scales and $\eta = 0.5$ setting at the higher

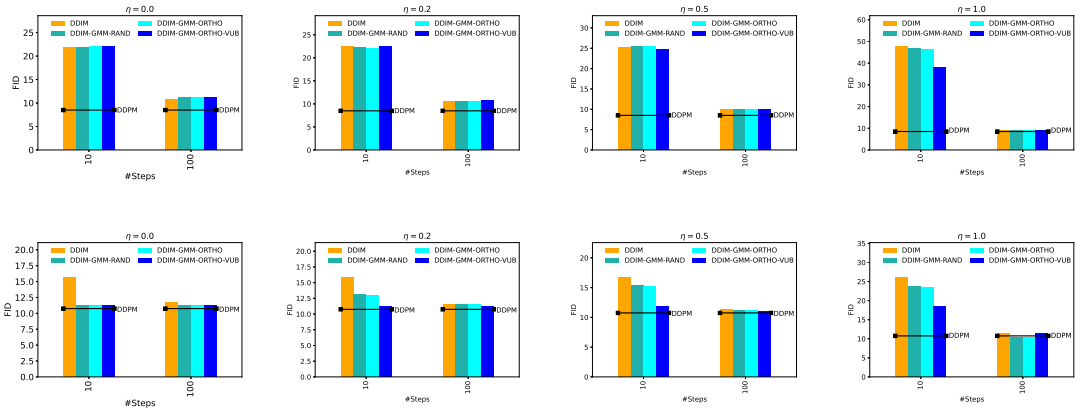

Figure 3: **Class-conditional ImageNet with Classifier Guidance**. FID (↓) with guidance scale of 1.0 (top) and 10.0 (bottom) respectively. The horizontal line is the DDPM baseline run for 1000 steps.

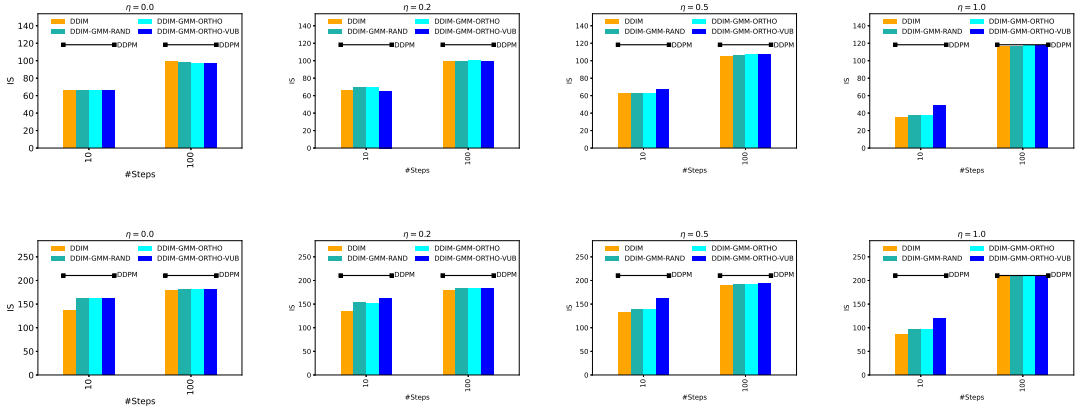

Figure 4: **Class-conditional ImageNet with Classifier Guidance**. IS (↑) with guidance scale of 1.0 (top) and 10.0 (bottom) respectively. The horizontal line is the DDPM baseline run for 1000 steps.

guidance scale (5). DDIM-GMM-ORTHO-VUB samples have consistently higher IS values under all settings. DDIM-GMM-ORTHO and DDIM-GMM-RAND yield the best IS (around 355) at $\eta = 1$ and DDIM-GMM-ORTHO-VUB yields the best FID (9.13) at $\eta = 0$. We posit that the similar performance of DDIM anf DDIM-GMM-* samplers using larger number of steps is due to the possibility that the multimodality of the true denoiser conditional distribution ($q(x_{t-1}|x_t)$) is modeled equally well by both the samplers (Guo et al., 2023; Xiao et al., 2022). Appendix A.14-A.15 show some qualitative results of sampling with the proposed approach compared to original DDIM.

4.3 Text-to-Image Generation

In this section, we experiment with a pretrained text-to-image diffusion model. Specifically, we use the publicly available Stable Diffusion v2.1 (Rombach et al., 2022) on a subset of 30,000 text and image pairs from the large scale COYO-700M image-text pair dataset (Byeon et al., 2022). Stable Diffusion v2.1 is a text-to-image diffusion model conditioned on text captions. It is trained on a subset of the large-scale LAION-5B image-text pair dataset (Schuhmann et al., 2022). We use DDIM and the variants of DDIM-GMM samplers, each with 5 and 10 sampling steps, to generate images at 256x256 resolution conditioned on captions from the COYO-700M data subset. Fig. 7

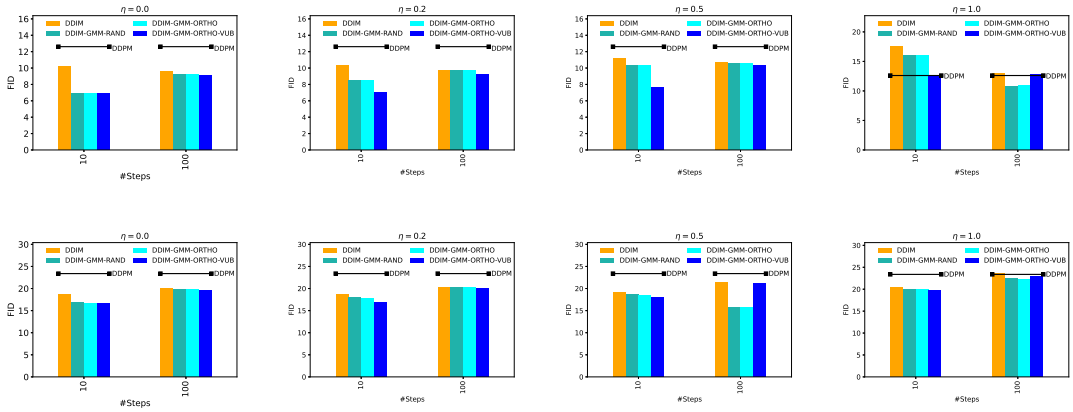

Figure 5: **Class-conditional ImageNet with Classifier-free Guidance**. FID (↓) with guidance scale of 2.5 (top) and 5.0 (bottom) respectively. The horizontal line is the DDPM baseline run for 1000 steps.

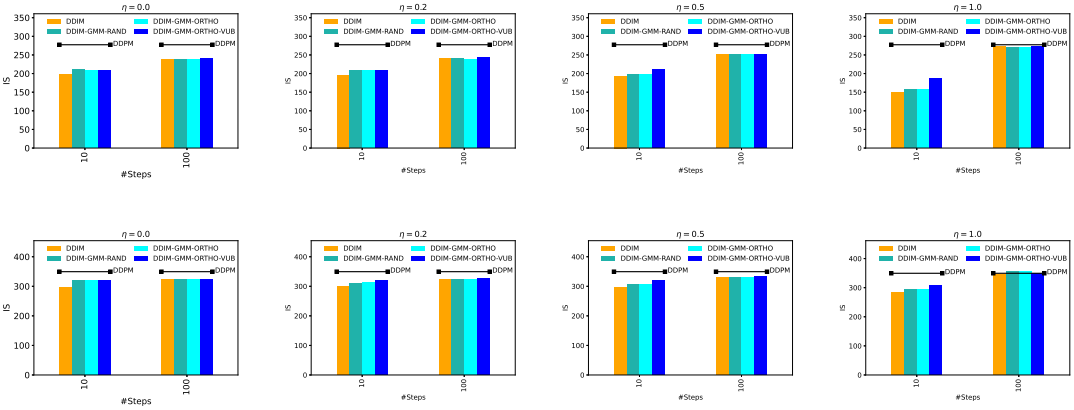

Figure 6: **Class-conditional ImageNet with Classifier-free Guidance**. IS (↑) with guidance scale of 2.5 (top) and 5.0 (bottom) respectively. The horizontal line is the DDPM baseline run for 1000 steps.

and 8 show the FID and IS metrics respectively. The FID metric improves consistently with the DDIM-GMM-* samplers relative to DDIM for all settings of η using 10 sampling steps. The relative improvements with DDIM-GMM-ORTHO-VUB over DDIM are more significant with increasing η compared to DDIM-GMM-RAND and DDIM-GMM-ORTHO suggesting better exploration of latent space, similar to results with unconditional models. Using 5 sampling steps, the DDIM-GMM-* samplers show most improvements with $\eta = 1$. On the IS metric, all variants of DDIM-GMM samplers show significant improvements over DDIM under different settings of η and sampling steps.

4.4 Sharing GMM parameters across sampling steps

As discussed in Section 3.1.3, we also experiment with sharing GMM parameters $\mathcal{M}_t$ across sampling steps t by choosing the offsets only once followed by orthogonalization (SVD), scaling and variance upper bounding. This saves some compute time during initialization. Table 1 compares the FIDs between DDIM-GMM-ORTHO-VUB samplers that share GMM parameters (ORTHO-VUB*) with the corresponding ones that set them independently (ORTHO-VUB) across sampling steps t. We observe that there is no significant (more than 1 FID point) impact on sample FIDs across all

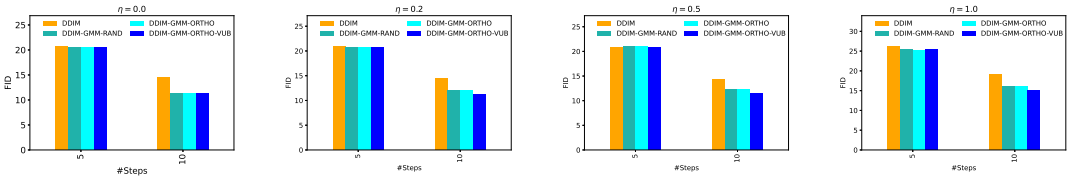

Figure 7: **Text-to-Image Generation**. FID (↓) on a 30k subset of COYO-700M using the Stable Diffusion v2.1 model. Classifier-free guidance with a scale of 7.5 is used during inference.

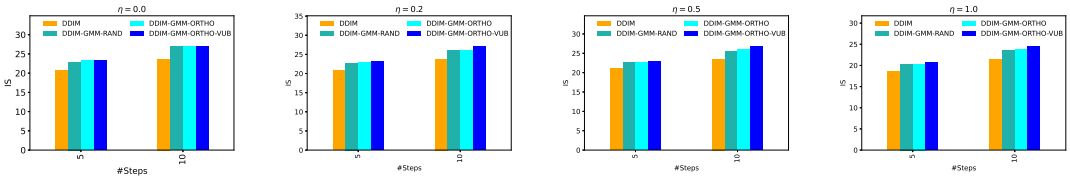

Figure 8: **Text-to-Image Generation**. IS (↑) on a 30k subset of COYO-700M using the Stable Diffusion v2.1 model. Classifier-free guidance with a scale of 7.5 is used during inference.

datasets, except with FFHQ with 10 sampling steps using $\eta = 1$, where the shared parameter sampler yields slightly better results. Here we show results with unconditional models on CelebAHQ and FFHQ and class-conditional model on ImageNet without any guidance (Appendix A.8). More results can be found in Tables 5 to 12 in the Appendix. Appendix A.9 discusses ablations on number of mixture components and offset scale s.

Table 1: **Sharing GMM Parameters across sampling steps**. FID (↓)

	Dataset	CelebAHQ		FFHQ		ImageNet	
	Steps	10	100	10	100	10	100
η							
0	ORTHO-VUB	27.94	11.44	26.46	11.12	36.35	19.81
0	ORTHO-VUB*	27.84	11.42	27.18	11.24	36.27	19.70
1.0	ORTHO-VUB	60.65	15.60	69.25	11.44	62.71	16.75
1.0	ORTHO-VUB*	61.15	15.94	**67.72**	11.38	63.02	16.88

5 Conclusions

We propose improved DDIM sampling by using Gaussian mixture transition kernels whose marginal first and second order moments match the corresponding moments of the DDPM forward marginals. Our experiments suggest that moment matching is sufficient to produce samples of the same or better quality than the original DDIM sampler. This is especially true if the number of sampling steps is small (e.g. 10) using unconditional models trained on CelebAHQ and FFHQ. For classifier guided ImageNet class-conditional models, at higher guidance weight (10), the GMM kernel based samplers lead to improvements in both FID and IS metrics under almost all settings of η and number of sampling steps (10 and 100). Similar improvements are seen with classifier-free guidance, especially with fewer sampling steps (10). This seems to suggest that the GMM kernel allows for a better exploration of the latent space with a small number of sampling steps. The gap between DDIM and the proposed DDIM-GMM becomes smaller with larger number of steps using training-free GMM parameter selection. We also demonstrate that DDIM-GMM shows improvements over DDIM in text-to-image generation with fewer sampling steps. An interesting future direction would be to optimize the GMM parameters $\mathcal{M}_t$ to maximize a suitable metric such as KID (Watson et al., 2021) or FID (Mathiasen & Hvilshøj, 2021) on the training dataset.

REFERENCES

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

Minh Do. Fast approximation of kullback-leibler distance for dependence trees and hidden markov models. *Signal Processing Letters, IEEE*, 10:115 – 118, 05 2003.

Gene H. Golub. Some modified matrix eigenvalue problems. *SIAM Review*, 15(2):318–334, 1973.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Hanzhong Allan Guo, Cheng Lu, Fan Bao, Tianyu Pang, Shuicheng YAN, Chao Du, and Chongxuan Li. Gaussian mixture solvers for diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

John Hershey and Peder Olsen. Approximating the kullback leibler divergence between gaussian mixture models. volume 4, pp. IV–317, 05 2007.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022.

Alexia Jolicoeur-Martineau, Ke Li, Rémi Piché-Taillefer, Tal Kachman, and Ioannis Mitliagkas. Gotta go fast when generating data with score-based models. *arXiv preprint arXiv:2105.14080*, 2021.

I.T. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 4401–4410. Computer Vision Foundation / IEEE, 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Proc. NeurIPS*, 2022.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2023.

Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. 2021.

Calvin Luo. Understanding diffusion models: A unified perspective. *ArXiv*, abs/2208.11970, 2022.

Alexander Mathiasen and Frederik Hvilshøj. Backpropagating through fréchet inception distance. *arXiv preprint arXiv:2009.14075*, 2021.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14297–14306, June 2023.

Eliya Nachmani, Robin San Roman, and Lior Wolf. Non gaussian denoising diffusion models. *arXiv preprint arXiv:2106.07582*, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 18–24 Jul 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1530–1538. PMLR, 2015.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1278–1286. PMLR, 2014.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Robin San-Roman, Eliya Nachmani, and Lior Wolf. Noise estimation for generative diffusion models. *arXiv preprint arXiv:2104.02600*, 2021.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265. PMLR, 07–09 Jul 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=St1giarCHLP`.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yang Song, Jascha Narain Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ArXiv*, abs/2011.13456, 2020.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *9th ISCA Speech Synthesis Workshop*, 2016a.

Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 1747–1756, 2016b.

Daniel Watson, William Chan, Jonathan Ho, and Mohammad Norouzi. Learning fast samplers for diffusion models by differentiating through sample quality. In *International Conference on Learning Representations*, 2021.

Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *ArXiv*, abs/2210.04628, 2022.

Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion GANs. In *International Conference on Learning Representations*, 2022.

Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *International Conference on Learning Representations*, 2023.

Qinsheng Zhang, Molei Tao, and Yongxin Chen. gddim: Generalized denoising diffusion implicit models, 2022.

A APPENDIX

A.1 RELATED WORK

Prior and concurrent work on accelerated sampling for pretrained diffusion models can be broadly categorized into implicit modeling (Song et al., 2021; Zhang et al., 2022; Watson et al., 2021), distillation (Luhman & Luhman, 2021; Salimans & Ho, 2022; Meng et al., 2023) and ODE solver (Song et al., 2020; Jolicoeur-Martineau et al., 2021; Zhang & Chen, 2023; Karras et al., 2022; Lu et al., 2023; Liu et al., 2022) based approaches. Watson et al. (2021) extend DDIMs by introducing a more general family of implicit distributions with learnable parameters trained with backpropagation using perceptual loss. While the proposed approach also introduces learnable parameters, our marginals are Gaussian mixtures and we ensure that the moments are matched exactly with those of the DDPM marginals. Zhang et al. (2022) analyze the workings of DDIM using a limiting case of Dirac distribution in the data space and generalize it to non-isotropic diffusion models. Other approaches propose accelerated sampling by modeling DDPMs with non-Gaussian noise (Nachmani et al., 2021) or learning noise levels of the reverse process separately (San-Roman et al., 2021). Different from these, the proposed approach introduces a different sampling kernel in the reverse process of the DDIM framework (Song et al., 2021). Our work is also complementary to distillation based approaches, which might further benefit from an improved DDIM teacher (Salimans & Ho, 2022; Meng et al., 2023).

By treating sampling as solving reverse direction diffusion ODEs (Song et al., 2020), acceleration is achieved by discretization with linear (Song et al., 2021) or higher order approximations (Jolicoeur-Martineau et al., 2021; Lu et al., 2022; Zhang & Chen, 2023). Being a moment matching version of DDIM, the proposed approach can be thought of as a linear solver. It is observed that higher-order solvers are inherently unstable in the guided sampling regime, especially if the guidance weight is high (Lu et al., 2023). Our empirical results suggest that, even with a high guidance weight, moment matching based DDIM-GMM is beneficial for guided sampling with few sampling steps.

A.2 EXPERIMENTAL DETAILS

We provide additional details on the experiments reported in Section 4, specifically for the CelebAHQ, FFHQ and ImageNet experiments. All our diffusion models are trained in the latent space of a VQVAE (Rombach et al., 2022). The input images to the VQVAE are at a resolution of 256x256 pixels. Each of the VQVAEs are trained on a large scale dataset. Specifically the VQVAEs for unconditional generation on CelebAHQ and class-conditional generation on ImageNet are trained on OpenImages. We use the publicly available $f4$ VQVAE (Table 8, Section D.2. of Rombach et al. (2022)) for training CelebAHQ models and $f8$ VQVAE for class-conditional ImageNet models respectively.The *f4* VQVAE (#embeddings=8192) does not use attention layers at any resolution within the model architecture, whereas the $f8$ VQVAE (#embeddings=16384) uses attention at resolution 32. We train a $f4$ VQVAE (#embeddings=8192), with no attention layers, on ImageNet for 712k steps and use its latent space to train the diffusion models on FFHQ.

All our diffusion models are trained with 1000 forward steps using a linear noise ($\beta_t = 1 - \frac{\alpha_t}{\alpha_{t-1}}$) schedule of $[\beta_0 = 0.0015, \beta_{1000} = 0.0195]$. We use the U-Net architecture (Ho et al., 2020; Rombach et al., 2022) for the denoiser. Specifically, the unconditional U-Net encoders operating on the $f4$ VQVAE latent space have four 2x downsampling levels with channel multiplication factors of $[1, 2, 3, 4]$ starting from a base set of 224 channels. Each level uses two residual blocks. Attention blocks are used within levels at downsampling factors $[2, 4, 8]$. Similarly, the class-conditional U-Net encoders operating on $f8$ VQVAE latent space have three 2x downsampling levels with channel multiplication factors of $[1, 2, 4]$ starting from a base set of 256 channels. Other architectural details remain the same as before except that the attention blocks are at downsampling levels [1, 2, 4].

For each DDIM-GMM variant, we choose a GMM with 8 mixture components with uniform priors (π_t^k=0.125) for all steps t. We also search for the best value of scaling s among $\{0.01, 0.1, 1.0, 10.0\}$ and report the best result with the chosen value for s. It is possible that for some choices of s, the diagonal elements of $\mathbf{\Delta}_k^t$ or the corresponding upper bounds in the DDIM-GMM-VUB sampler could be larger than σ_t^2. In such cases we clip the negative elements of $(\sigma_t^2 \boldsymbol{I} - diag\_approx(\mathbf{\Delta}_t^k))$ to zero, which amounts to sampling with zero variances in those dimensions in the latent space.

A.3 Proof of Constraints on GMM Parameters

Our proof for the constraints in Eq. 8 follows by induction (Song et al., 2021). The marginal of $\boldsymbol{x}_T$ is already equal to the DDPM marginal at step T by definition (Eq. 5). We show below that the marginals of all the random variables $\boldsymbol{x}_t, t < T$ are Gaussian mixtures with their first and second order moments equal to the desired values, given the constraints in Eq. 8. We derive the forms of the marginals for $T-1$ and $T-2$ and the proof follows inductively for all $t < T-2$. Using Bayes' rule, the marginal at $\boldsymbol{x}_{T-1}$ is given by

$$
\begin{aligned}
q_{\sigma,\mathcal{M}}(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0) &= \int_{\boldsymbol{x}_T} q_{\sigma,\mathcal{M}}(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T,\boldsymbol{x}_0) q_{\sigma,\mathcal{M}}(\boldsymbol{x}_T|\boldsymbol{x}_0)\,\mathrm{d}\boldsymbol{x}_T \\
&= \int_{\boldsymbol{x}_T} \sum_{k=1}^{K} \pi_T^k \mathcal{N}\left(\sqrt{\alpha_{T-1}}\boldsymbol{x}_0 + \sqrt{1-\alpha_{T-1}-\sigma_T^2}.\frac{\boldsymbol{x}_T - \sqrt{\alpha_T}\boldsymbol{x}_0}{\sqrt{1-\alpha_T}} + \boldsymbol{\delta}_T^k, \sigma_T^2\boldsymbol{I} - \boldsymbol{\Delta}_T^k\right) \\
&\qquad\qquad q_{\sigma,\mathcal{M}}(\boldsymbol{x}_T|\boldsymbol{x}_0)\,\mathrm{d}\boldsymbol{x}_T \\
&= \sum_{k=1}^{K} \pi_T^k \int_{\boldsymbol{x}_T} \mathcal{N}\left(\sqrt{\alpha_{T-1}}\boldsymbol{x}_0 + \sqrt{1-\alpha_{T-1}-\sigma_T^2}.\frac{\boldsymbol{x}_T - \sqrt{\alpha_T}\boldsymbol{x}_0}{\sqrt{1-\alpha_T}} + \boldsymbol{\delta}_T^k, \sigma_T^2\boldsymbol{I} - \boldsymbol{\Delta}_T^k\right) \\
&\qquad\qquad \mathcal{N}\left(\sqrt{\alpha_T}\boldsymbol{x}_0, (1-\alpha_T)\boldsymbol{I}\right)\,\mathrm{d}\boldsymbol{x}_T, \\
&= \sum_{k=1}^{K} \pi_T^k \mathcal{N}\left(\sqrt{\alpha_{T-1}}\boldsymbol{x}_0 + \boldsymbol{\delta}_T^k, (1-\alpha_{T-1})\boldsymbol{I} - \boldsymbol{\Delta}_T^k\right), \qquad (13)
\end{aligned}
$$

which is also a GMM with the same mixing weights π_T^k. This is due to the fact that each of the above integrals is a Gaussian, whose parameters can be determined by using Gaussian marginalization identities (Bishop, 2006)(2.115). The mean $\boldsymbol{\mu}_{T-1}^{GMM}$ and the covariance $\boldsymbol{\Sigma}_{T-1}^{GMM}$ parameters of the above GMM are given by

$$
\begin{aligned}
\boldsymbol{\mu}_{T-1}^{GMM} &= \sum_{k=1}^{K} \pi_T^k \left(\sqrt{\alpha_{T-1}}\boldsymbol{x}_0 + \boldsymbol{\delta}_T^k\right) \\
&= \sqrt{\alpha_{T-1}}\boldsymbol{x}_0 + \sum_{k=1}^{K} \pi_T^k \boldsymbol{\delta}_T^k \\
\boldsymbol{\Sigma}_{T-1}^{GMM} &= \sum_{k=1}^{K} \pi_T^k \left((1-\alpha_{T-1})\boldsymbol{I} - \boldsymbol{\Delta}_T^k\right) + \sum_{k=1}^{K} \pi_T^k (\boldsymbol{\delta}_T^k - \bar{\boldsymbol{\delta}}_T)(\boldsymbol{\delta}_T^k - \bar{\boldsymbol{\delta}}_T)^T \\
&= (1-\alpha_{T-1})\boldsymbol{I} + \sum_{k=1}^{K} \pi_T^k \left((\boldsymbol{\delta}_T^k - \bar{\boldsymbol{\delta}}_T)(\boldsymbol{\delta}_T^k - \bar{\boldsymbol{\delta}}_T)^T - \boldsymbol{\Delta}_T^k\right), \qquad (14)
\end{aligned}
$$

where $\bar{\boldsymbol{\delta}}_T = \sum_{k=1}^{K} \pi_T^k \boldsymbol{\delta}_T^k$. It is straightforward to verify that these are equal to the desired means and covariance parameters of the equivalent DDPM forward marginal if the constraints in Eq. 8 are satisfied. Specifically the constraint $\bar{\boldsymbol{\delta}}_T = 0$ and either one of the constraints on $\boldsymbol{\Delta}_T^k$ in Eq. 8 lead to the following expressions for the first and second order moments:

$$
\begin{aligned}
\boldsymbol{\mu}_{T-1}^{GMM} &= \sqrt{\alpha_{T-1}}\boldsymbol{x}_0 \\
\boldsymbol{\Sigma}_{T-1}^{GMM} &= (1-\alpha_{T-1})\boldsymbol{I}, \qquad (15)
\end{aligned}
$$

as desired.

The marginal of $\boldsymbol{x}_{T-2}$ can be derived similarly by invoking Bayes' rule and using the form of the GMM for $\boldsymbol{x}_{T-1}$. Specifically

$$
\begin{aligned}
q_{\sigma,\mathcal{M}}(\boldsymbol{x}_{T-2}|\boldsymbol{x}_0) &= \int_{\boldsymbol{x}_{T-1}} q_{\sigma,\mathcal{M}}(\boldsymbol{x}_{T-2}|\boldsymbol{x}_{T-1},\boldsymbol{x}_0) q_{\sigma,\mathcal{M}}(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0)\,\mathrm{d}\boldsymbol{x}_{T-1} \\
&= \int_{\boldsymbol{x}_{T-1}} \sum_{l=1}^{L} \pi_{T-1}^{l} \mathcal{N}\left(\sqrt{\alpha_{T-2}}\boldsymbol{x}_0 + \sqrt{1-\alpha_{T-2}-\sigma_{T-1}^2}.\frac{\boldsymbol{x}_{T-1}-\sqrt{\alpha_{T-1}}\boldsymbol{x}_0}{\sqrt{1-\alpha_{T-1}}} + \boldsymbol{\delta}_{T-1}^{l}, \sigma_{T-1}^2\boldsymbol{I} - \boldsymbol{\Delta}_{T-1}^{l}\right) \\
&\qquad\qquad q_{\sigma,\mathcal{M}}(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0)\,\mathrm{d}\boldsymbol{x}_{T-1} \\
&= \int_{\boldsymbol{x}_{T-1}} \left\{\sum_{l=1}^{L} \pi_{T-1}^{l} \mathcal{N}\left(\sqrt{\alpha_{T-2}}\boldsymbol{x}_0 + \sqrt{1-\alpha_{T-2}-\sigma_{T-1}^2}.\frac{\boldsymbol{x}_{T-1}-\sqrt{\alpha_{T-1}}\boldsymbol{x}_0}{\sqrt{1-\alpha_{T-1}}} + \boldsymbol{\delta}_{T-1}^{l}, \sigma_{T-1}^2\boldsymbol{I} - \boldsymbol{\Delta}_{T-1}^{l}\right)\right\} \\
&\qquad\qquad \left\{\sum_{k=1}^{K} \pi_{T}^{k} \mathcal{N}\left(\sqrt{\alpha_{T-1}}\boldsymbol{x}_0 + \boldsymbol{\delta}_{T}^{k}, (1-\alpha_{T-1})\boldsymbol{I} - \boldsymbol{\Delta}_{T}^{k}\right)\right\}\,\mathrm{d}\boldsymbol{x}_{T-1} \\
&= \sum_{k=1}^{K}\sum_{l=1}^{L} \pi_{T}^{k}\pi_{T-1}^{l} \int_{\boldsymbol{x}_{T-1}} \mathcal{N}\left(\sqrt{\alpha_{T-2}}\boldsymbol{x}_0 + \sqrt{1-\alpha_{T-2}-\sigma_{T-1}^2}.\frac{\boldsymbol{x}_{T-1}-\sqrt{\alpha_{T-1}}\boldsymbol{x}_0}{\sqrt{1-\alpha_{T-1}}} + \boldsymbol{\delta}_{T-1}^{l}, \sigma_{T-1}^2\boldsymbol{I} - \boldsymbol{\Delta}_{T-1}^{l}\right) \\
&\qquad\qquad \mathcal{N}\left(\sqrt{\alpha_{T-1}}\boldsymbol{x}_0 + \boldsymbol{\delta}_{T}^{k}, (1-\alpha_{T-1})\boldsymbol{I} - \boldsymbol{\Delta}_{T}^{k}\right)\,\mathrm{d}\boldsymbol{x}_{T-1} \\
&= \sum_{k=1}^{K}\sum_{l=1}^{L} \pi_{T}^{k}\pi_{T-1}^{l} \mathcal{N}\left(\boldsymbol{\mu}_{k,l}^{T,T-1}, \boldsymbol{\Sigma}_{k,l}^{T,T-1}\right), \qquad (16)
\end{aligned}
$$

where we assume that the transition kernel from $\boldsymbol{x}_{T-1}$ to $\boldsymbol{x}_{T-2}$ is a GMM with L components and parameters $\left(\pi_{T-1}^{l}, \boldsymbol{\delta}_{T-1}^{l}, \boldsymbol{\Delta}_{T-1}^{l}\right), l = 1 \ldots L$. Each of the integrals above is a Gaussian whose mean $\boldsymbol{\mu}_{k,l}^{T,T-1}$ and covariance $\boldsymbol{\Sigma}_{k,l}^{T,T-1}$ parameters can be deduced by invoking Gaussian marginalization identities (Bishop, 2006)(2.115) and are given by:

$$
\begin{aligned}
\boldsymbol{\mu}_{k,l}^{T,T-1} &= \sqrt{\alpha_{T-2}}\boldsymbol{x}_0 + \frac{\sqrt{1-\alpha_{T-2}-\sigma_{T-1}^2}}{\sqrt{1-\alpha_{T-1}}}\boldsymbol{\delta}_{T}^{k} + \boldsymbol{\delta}_{T-1}^{l} \\
\boldsymbol{\Sigma}_{k,l}^{T,T-1} &= (1-\alpha_{T-2})\boldsymbol{I} - \frac{1-\alpha_{T-2}-\sigma_{T-1}^2}{1-\alpha_{T-1}}\boldsymbol{\Delta}_{T}^{k} - \boldsymbol{\Delta}_{T-1}^{l}. \qquad (17)
\end{aligned}
$$

Using the above expressions for the mean and covariance parameters of individual Gaussian components, the corresponding parameters $\boldsymbol{\mu}_{T-2}^{GMM}$ and $\boldsymbol{\Sigma}_{T-2}^{GMM}$ for the GMM marginal of $\boldsymbol{x}_{T-2}$ are given by:

$$
\begin{aligned}
\boldsymbol{\mu}_{T-2}^{GMM} &= \sum_{k=1}^{K}\sum_{l=1}^{L} \pi_{T}^{k}\pi_{T-1}^{l}\boldsymbol{\mu}_{k,l}^{T,T-1} \\
&= \sqrt{\alpha_{T-2}}\boldsymbol{x}_0 \qquad (18) \\
& \qquad (19)
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{\Sigma}_{T-2}^{GMM} &= \sum_{k=1}^{K}\sum_{l=1}^{L} \pi_{T}^{k}\pi_{T-1}^{l}\boldsymbol{\Sigma}_{k,l}^{T,T-1} + \sum_{k=1}^{K}\sum_{l=1}^{L} \pi_{T}^{k}\pi_{T-1}^{l}[A\boldsymbol{\delta}_{k}^{T} + \boldsymbol{\delta}_{l}^{T-1}][A\boldsymbol{\delta}_{k}^{T} + \boldsymbol{\delta}_{l}^{T-1}]^{T} \\
&= (1-\alpha_{T-2})\boldsymbol{I}, \qquad (20)
\end{aligned}
$$

where

$$A = \frac{\sqrt{1 - \alpha_{T-2} - \sigma_{T-1}^2}}{\sqrt{1 - \alpha_{T-1}}}. \tag{21}$$

We have made use of the constraints in Eq. 8, for the parameters $(\pi_k^T, \pi_l^{T-1}, \boldsymbol{\delta}_k^T, \boldsymbol{\delta}_l^{T-1}, \boldsymbol{\Delta}_k^T, \boldsymbol{\Delta}_l^{T-1})$, to arrive at the above expressions and noting that A is independent of the GMM parameters $\mathcal{M}_T$ and $\mathcal{M}_{T-1}$. The first and second order moments in Eq. 19 and Eq. 20 correspond to the DDPM forward marginal moments of $\boldsymbol{x}_{T-2}$. The proof for all latents $\boldsymbol{x}_t, t < T-2$ follows from a similar argument as above noting that the form of the marginal in Eq. 16 is a GMM with $M = KL$ components. Further each component's mean and covariances in Eq. 17 carry future ($T-1$ and T) step transition kernels' offset parameters ($\boldsymbol{\delta}$'s and $\boldsymbol{\Delta}$'s) as linear additive factors with coefficients (A, A^2 and 1) that are independent of those parameters. This makes it easier to see why proof by induction should work for $t < T-2$.

A.4 Variance Upper Bounds

The upper bounds of the eigenvalues of the matrix $\boldsymbol{\Delta}_t^k$ are tractable because the matrix is a weighted sum of outer-products of mean centered orthonormal vectors. Specifically $\boldsymbol{\Delta}_t^k$ can be written as

$$\boldsymbol{\Delta}_t^k = \frac{s^2}{K\pi_t^k} \left(\sum_{l=1}^{K} \pi_t^k (\boldsymbol{u}_t^l)(\boldsymbol{u}_t^l)^T - \bar{\boldsymbol{u}}_t \bar{\boldsymbol{u}}_t^T \right), \tag{22}$$

where $\boldsymbol{u}_t^k = \boldsymbol{U}_t[k]$ (Section 3.1.2). Diagonalizing the first term in Eq. 22 by pre and post multiplying by the matrix $\boldsymbol{U}_t[1:K]$ leads to a matrix $\boldsymbol{M}_t^k$, which is a sum of a diagonal matrix and a rank one matrix,

$$\begin{aligned} \boldsymbol{M}_t^k &= \boldsymbol{U}_t[1:K]^T \boldsymbol{\Delta}_t^k \boldsymbol{U}_t[1:K] \\ &= \frac{s^2}{K\pi_t^k} (\boldsymbol{D}_t^k - \boldsymbol{\pi}_t \boldsymbol{\pi}_t^T), \end{aligned} \tag{23}$$

where $\boldsymbol{D}_t^k$ is a diagonal matrix with $\pi_t^k, k = 1 \dots K$, along its diagonal and $\boldsymbol{\pi}_t$ is a column vector of mixture proportions π_t^k. Using a bound (Golub, 1973) on the eigenvalues of a diagonal matrix modified by a rank one matrix, if $\lambda_i, i = 1 \dots K$ are the eigenvalues of $\boldsymbol{M}_t^k$, then

$$\begin{aligned} \frac{s^2}{K\pi_t^k} \left(\pi_t^1 - \sum_{l=1}^{K} (\pi_t^l)^2 \right) \leq \lambda_1 \leq \frac{s^2}{K\pi_t^k} \pi_t^1 \\ \frac{s^2}{K\pi_t^k} \pi_t^{i-1} \leq \lambda_i \leq \frac{s^2}{K\pi_t^k} \pi_t^i, \qquad i = 2 \dots K. \end{aligned} \tag{24}$$

A.5 Forward Process

We can derive the forward process using Bayes' rule. Specifically

$$q_{\sigma,\mathcal{M}}(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}, \boldsymbol{x}_0) = \frac{q_{\sigma,\mathcal{M}}(\boldsymbol{x}_{t-1} | \boldsymbol{x}_t, \boldsymbol{x}_0) q_{\sigma,\mathcal{M}}(\boldsymbol{x}_t | \boldsymbol{x}_0)}{q_{\sigma,\mathcal{M}}(\boldsymbol{x}_{t-1} | \boldsymbol{x}_0)} \qquad = \frac{GMM(t, t-1) GMM(t)}{GMM(t-1)}$$

where $GMM(t, t-1)$ is the transition GMM from step t to $t-1$ and $GMM(t-1)$ and $GMM(t)$ are the marginal GMMs at steps $t-1$ and t respectively. Note that the inference process of the proposed implicit model is non-Gaussian and non-Markovian in general and different from Gaussian diffusion.

A.6 Upper bound of ELBO using the DDIM-GMM inference process

In this section we provide an upper bound for the ELBO loss using the proposed DDIM-GMM inference process in terms of an augmented version of the DDPM $\mathcal{L}_{simple,w}$ loss, w.r.t. the denoiser

parameters θ. The ELBO loss $\mathcal{L}_{ELBO,q_{\sigma,\mathcal{M}}}(\theta)$ using the proposed DDIM-GMM inference process is given by

$$\begin{aligned}\mathcal{L}_{ELBO,q_{\sigma,\mathcal{M}}}(\theta) &= \mathbb{E}_{q_{\sigma,\mathcal{M}}}\left[D_{KL}(q_{\sigma,\mathcal{M}}(\boldsymbol{x}_T|\boldsymbol{x}_0)||p_\theta(\boldsymbol{x}_T)) + \sum_{t=2}^{T} D_{KL}(q_{\sigma,\mathcal{M}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)||p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right] \\ &\quad - \mathbb{E}_{q_{\sigma,\mathcal{M}}}\left[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] \\ &= \mathbb{E}_{q_{\sigma,\mathcal{M}}}\left[\sum_{t=2}^{T} D_{KL}(q_{\sigma,\mathcal{M}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)||p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) - \log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] + const.,\end{aligned} \tag{25}$$

where $const.$ is a term independent of θ because $p_\theta(\boldsymbol{x}_T) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. We ignore the constant term and assume a normal likelihood function for the observation $\boldsymbol{x}_0$ given $\boldsymbol{x}_1$, i.e.,

$$p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1) = \mathcal{N}(\boldsymbol{f}_\theta(\boldsymbol{x}_1, 1), \sigma_1^2\boldsymbol{I}), \tag{26}$$

where $\boldsymbol{f}_\theta(\boldsymbol{x}_t, t)$ is the denoiser estimate of $\boldsymbol{x}_0$, given by:

$$\boldsymbol{f}_\theta(\boldsymbol{x}_t, t) = \frac{\boldsymbol{x}_t - \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)}{\sqrt{\alpha_t}}. \tag{27}$$

The ELBO loss in Eq. 25 reduces to

$$\begin{aligned}\mathcal{L}_{ELBO,q_{\sigma,\mathcal{M}}}(\theta) =& \sum_{t=2}^{T} \mathbb{E}_{q_{\sigma,\mathcal{M}_t}}\left[D_{KL}(q_{\sigma,\mathcal{M}_t}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)||q_{\sigma,\mathcal{M}_t}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{f}_\theta(\boldsymbol{x}_t,t)))\right] \\ &+ \mathbb{E}_{q_{\sigma,\mathcal{M}}(\boldsymbol{x}_0,\boldsymbol{x}_1)}\left[-\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] \\ :=& K_1 + K_2 \\ =& \sum_{t=2}^{T} K_{1,t} + K_2,\end{aligned} \tag{28}$$

where we use $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = q_{\sigma,\mathcal{M}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{f}_\theta(\boldsymbol{x}_t, t))$ as in DDIM (Song et al., 2021). The first term K_1 involves KL-Divergences between mixtures of Gaussians, which is analytically intractable. However, we can use a suitable upper bound (Hershey & Olsen, 2007) as a surrogate for optimization. Assuming that there is a one-to-one correspondence between the mixture components of the GMMs using the true and estimated value of $\boldsymbol{x}_0$ above, we can use the matched bound (Hershey & Olsen, 2007; Do, 2003) as the upper bound to each of the KLD terms $K_{1,t}$ at step t, i.e. for any $t > 1$

$$\begin{aligned}K_{1,t} &\leq \mathbb{E}_{q_{\sigma,\mathcal{M}}}\left[\sum_k \pi_t^k D_{KL}((q_{\sigma,\mathcal{M}_t^k}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)||q_{\sigma,\mathcal{M}_t^k}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{f}_\theta(\boldsymbol{x}_t,t)))))\right] \\ &\leq \mathbb{E}_{q(\boldsymbol{x}_0)q_{\sigma,\mathcal{M}}(\boldsymbol{x}_t|\boldsymbol{x}_0)\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\left(\sum_k \frac{\pi_t^k}{\nu_t^k}\right)\frac{(1-\alpha_t)}{2\alpha_t}||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t)||^2\right] \\ &= \sum_l \xi_t^{GMM,l}\mathbb{E}_{q(\boldsymbol{x}_0)q_{\sigma,\mathcal{M}}^l(\boldsymbol{x}_t|\boldsymbol{x}_0)\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\left(\sum_k \frac{\pi_t^k}{\nu_t^k}\right)\frac{(1-\alpha_t)}{2\alpha_t}||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t)||^2\right] \\ &= \mathbb{E}_{l\sim\xi_t^{GMM}q(\boldsymbol{x}_0)q_{\sigma,\mathcal{M}}^l(\boldsymbol{x}_t|\boldsymbol{x}_0)\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\left(\sum_k \frac{\pi_t^k}{\nu_t^k}\right)\frac{(1-\alpha_t)}{2\alpha_t}||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t)||^2\right]\end{aligned} \tag{29}$$

where ν_t^k is the minimum variance within a diagonal approximation of the covariance matrix $\sigma_t^2\boldsymbol{I} - \boldsymbol{\Delta}_t^k$, i.e., $\nu_t^k = \min diag((\sigma_t^2\boldsymbol{I} - diag\_approx(\boldsymbol{\Delta}_t^k)))$. Note that in the above, $q_{\sigma,\mathcal{M}_t^k}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)$ is used to refer to the kth mixture component's density function of the GMM transition kernel. Similarly, $q_{\sigma,\mathcal{M}}^l(\boldsymbol{x}_t|\boldsymbol{x}_0)$ refers to the lth component of the DDIM-GMM's forward marginal GMM at step t. The upper bound of Eq. 29 can be interpreted as an augmented form of $\mathcal{L}_{simple,w}$ with weights

$$w_t = \left(\sum_k \frac{\pi_t^k}{\nu_t^k}\right)\frac{(1-\alpha_t)}{2\alpha_t} \tag{30}$$

and the DDPM marginals' mean and covariance randomly modified with shifts from one of the DDIM-GMM forward marginal's components at every step t (e.g., Eq. 17 for $t = T - 2$). The choice of the shifts is according to a discrete distribution with proportions given by the DDIM-GMM marginal's mixture priors ξ_t^{GMM} (e.g. Eq. 16 for $t = T - 2$).

For $t = 1$, the loss term K_2 is given by

$$\begin{aligned} K_2 &= \mathbb{E}_{q_{\sigma,\mathcal{M}}(\boldsymbol{x}_0,\boldsymbol{x}_1)}\left[-\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] \\ &= \mathbb{E}_{q(\boldsymbol{x}_0)q_{\sigma,\mathcal{M}}(\boldsymbol{x}_1|\boldsymbol{x}_0)\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\frac{(1-\alpha_1)}{2\sigma_1^2\alpha_1}||\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_1,1)||^2\right] + const. \\ &= \mathbb{E}_{l\sim\xi_1^{GMM}q(\boldsymbol{x}_0)q^l_{\sigma,\mathcal{M}}(\boldsymbol{x}_1|\boldsymbol{x}_0)\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\frac{(1-\alpha_1)}{2\sigma_1^2\alpha_1}||\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_1,1)||^2\right], \end{aligned} \tag{31}$$

where we have ignored the $const.$ term independent of θ. Combining Eqs. 29 and 31, the $\mathcal{L}_{ELBO,q_{\sigma,\mathcal{M}}}(\theta)$ can be interpreted as upper bounded by an augmented version of $\mathcal{L}_{simple,w}$ with weights w_t given in Eqs. 30 and 31.

A.7 DDIM-GMM AS A MULTIMODAL DENOISER

Recent works (Guo et al., 2023; Xiao et al., 2022) have shown that the target conditional distribution $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$, to be estimated by the denoiser $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$, is multimodal in real-world datasets. Here we show that the proposed DDIM-GMM sampling scheme addresses the unimodal assumption of the single Gaussian denoisers in pre-trained diffusion models better than DDIM. We start by showing that the proposed DDIM-GMM kernel yields a multimodal conditional distirbution $q_{\sigma,\mathcal{M}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$. Let the true data distribution $q(\boldsymbol{x}_0)$ be a Dirac distribution given by

$$q(\boldsymbol{x}_0) = \sum_i w_i \delta(\boldsymbol{x}_0 - \boldsymbol{x}_0^i), \tag{32}$$

where $\boldsymbol{x}_0^i$ are the observed data points. The DDIM-GMM denoiser's conditional distribution $q_{\sigma,\mathcal{M}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ can be obtained from Eq. 7 using Bayes' rule:

$$\begin{aligned} q_{\sigma,\mathcal{M}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) &= \int_{\boldsymbol{x}_0} q_{\sigma,\mathcal{M}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0) q_{\sigma,\mathcal{M}}(\boldsymbol{x}_0|\boldsymbol{x}_t)\,\mathrm{d}\boldsymbol{x}_0 \\ &\propto \int_{\boldsymbol{x}_0} q_{\sigma,\mathcal{M}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0) q_{\sigma,\mathcal{M}}(\boldsymbol{x}_t|\boldsymbol{x}_0) q(\boldsymbol{x}_0)\,\mathrm{d}\boldsymbol{x}_0 \\ &= \sum_i w_i q_{\sigma,\mathcal{M}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0^i) q_{\sigma,\mathcal{M}}(\boldsymbol{x}_t|\boldsymbol{x}_0^i), \end{aligned} \tag{33}$$

which is a mixture of Gaussians. This follows from the fact that $q_{\sigma,\mathcal{M}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0^i)$ is a mixture of Gaussians given by Eq. 7 and $q_{\sigma,\mathcal{M}}(\boldsymbol{x}_t|\boldsymbol{x}_0^i)$ is a scalar constant given $\boldsymbol{x}_t$. A similar argument holds even when the data distribution $q(\boldsymbol{x}_0)$ is a mixture of Gaussians. The resulting denoiser is a mixture of Gaussians, whose form can be obtained by noting that $q_{\sigma,\mathcal{M}}(\boldsymbol{x}_t|\boldsymbol{x}_0)$ is a GMM and accordingly completing squares within the integrand above (Bishop, 2006). A similar argument as above also enables DDIM sampler to model a multimodal denoising distribution $q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$. However the multimodality of the kernel $q_{\sigma,\mathcal{M}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0^i)$ in Eq. 33 enables DDIM-GMM to express more complex denoising distributions than DDIM. This has the potential to better match the unknown distribution $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$, especially when the number of sampling steps is small (Guo et al., 2023; Xiao et al., 2022), without any training or fine-tuning with specialized loss functions.

A.8 CLASS-CONDITIONAL IMAGENET WITHOUT ANY GUIDANCE

The FID and IS metrics of sampling from ImageNet class-conditional models without any guidance are shown in Fig. 9. From the results, we see that DDIM-GMM-ORTHO and DDIM-GMM-ORTHO-VUB sampling methods yield higher quality samples under the highest η setting. The FID is significantly better for the least number of sampling steps (10) (see Table 7, Appendix A.13.3),

whereas the IS is better for both the least (10) and the greatest number of sampling steps (100) (see Table 8, Appendix A.13.3). This can be attributed to a similar argument as for unconditional sampling, especially for the least number of sampling steps.

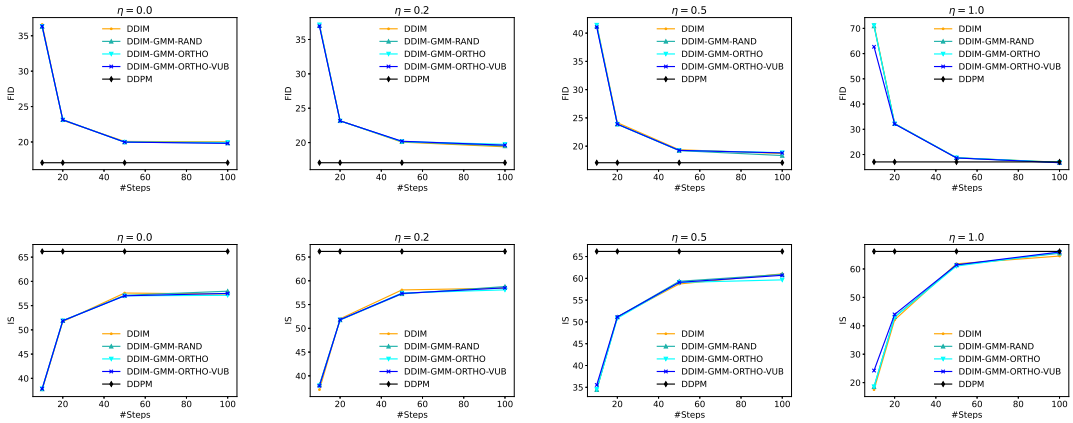

Figure 9: **Class-conditional ImageNet**. Top and bottom rows correspond to FID (↓) and IS (↑) metrics respectively. The horizontal line is the DDPM baseline run for 1000 steps.

A.9 Ablations

In this section, we perform an ablative study on the number of mixture components and offset scaling factor s of the GMM parameters using the unconditional model trained on the CelebAHQ dataset as described in Section 4. We use the GMM-ORTHO-VUB sampler and fix the mixture weights of the components to be uniform in all these experiments.

A.9.1 Number of mixture components

We compute the FID on the validation set by choosing one of 8, 256, or 1024 components (n) at each step during sampling, using different values of η. For each choice of n, we select the scale s among (0.01, 0.1, 1.0, 10.0) that leads to the lowest FID. The results are plotted in Fig. 10. For lower values of η (0, 0.2), the performance is the same across all the choices, since the offsets perturb only the means as the offset variances ($diag\_approx(\sigma_t^2 \boldsymbol{I} - \boldsymbol{\Delta}_t^k)$) are close to zero. At higher η values (0.5, 1.0), the offsets affect variances in as many dimensions and influence the exploration of latent spaces x_t. The choices 8 and 256 lead to better samples than 1024 because the latter restricts the variances in those many dimensions impacting exploration. 8 performs slightly better than 256 at $\eta = 0.5$ and vice-versa at $\eta = 1.0$. As the number of steps increases, all the choices lead to similar results, likely due to small s (See Table 5).

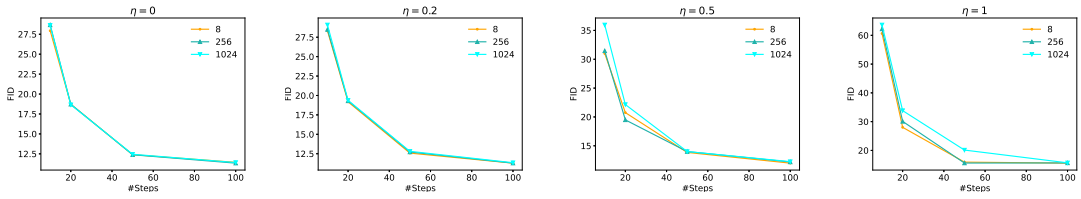

Figure 10: **CelebAHQ**. FID (↓). Ablations on the number of mixture components.

A.9.2 Offset scaling s

In order to study the effect of s, we fix the number of mixture components to 8 and choose a value for s within four choices: (0.01, 0.1, 1.0, 10.0). The results are shown in Fig. 11. The sample quality

is almost the same with smaller values of s (0.01-1.0). The highest value $s = 10$ gives the best results for the least number of sampling steps (10) (See Table 5). As the number of steps increases, this leads to poor quality samples and smaller offsets are preferable, with one exception: at $\eta = 1$ it is still the best choice for up to 50 steps. This can be explained using the hypothesis Guo et al. (2023); Xiao et al. (2022) that true denoising distributions are multimodal at fewer sampling steps and larger exploration (higher s) with a multimodal kernel is favorable. This advantage vanishes as the number of sampling steps increase. At the highest $\eta(= 1)$, we hypothesize that $s = 10$ reduces the offset variances ($diag\_approx(\sigma_t^2 \boldsymbol{I} - \boldsymbol{\Delta}_t^k)$) more than other choices, at least up to 50 steps, keeping sampling quality high.

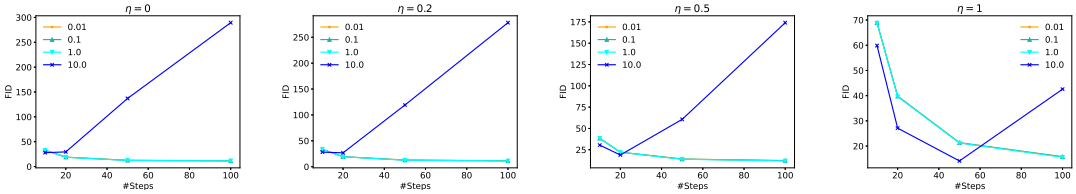

Figure 11: **CelebAHQ**. FID (↓). Ablations on the offset scaling factor s.

A.10 Computational Overhead

As discussed briefly in Section 3.1, the proposed approach introduces additional computational overhead in an initialization phase prior to sampling. All the GMM mean and variance offsets are precomputed and saved in memory before sampling. We can choose to precompute a single set of GMM parameters per batch or the entire sample set. We experimented with both the options and did not see a significant difference in metrics. So it is computationally more efficient to precompute offsets once and fix them. We also experimented with choosing different GMM parameters for different sampling steps t and found no significant difference with setting them the same across all t in our experiments. Due to storing the additional GMM parameters, there is some memory overhead relative to DDIM but it is negligible, especially in the scenario of choosing a single set of offsets across all samples and time steps. In the scenario of using different offsets across subsets (batches) of samples or sampling steps, the overhead scales linearly along the sample subset size and number of step dimensions. The dimensionality of the latent spaces $\boldsymbol{x}_t$ also influence the computational and memory requirements of the GMM offset parameters. For instance, it might be infeasible to compute the outer products of centered offsets (Eq. 11) if the dimensionality of the latent spaces is high, e.g. high-resolution image space diffusion models. In such cases, the DDIM-GMM-ORTHO-VUB sampler is more feasible as it provides an upper bound for the variance offsets without explicitly computing them.

A.11 Comparison with DPM-Solver

In this section, we compare DDIM-GMM-ORTHO-VUB and DPM-Solver (Lu et al., 2022) samplers on the class-conditional model trained on ImageNet. During inference, we use classifier-free guidance with weights (2.5, 5) and run each sampler for 10 and 100 steps. For the DDIM-GMM-ORTHO-VUB sampler, η is set to 0. The results listed in Table 4 suggest DDIM-GMM-ORTHO-VUB is superior to DPM-Solver in all cases, with the exception of lower guidance scale (2.5) using 10 sampling steps .

Steps	10		100	
Guidance Scale	2.5	5	2.5	5
DPM-Solver	9.75/**225.68**	19.22/309.14	9.30/239.11	19.82/323.87
DDIM-GMM-ORTHO-VUB	**6.94**/207.85	**16.61/319.13**	**9.13/240.88**	**19.66/323.84**

Table 2: Comparison with DPM-Solver. Class-conditional ImageNet with classifier free guidance (FID↓ / IS↑).

A.12 ADDITIONAL EXPERIMENTS

In this section, we report results on additional experiments on the LSUN benchmarks (Yu et al., 2015) using the same settings as the DDIM (Song et al., 2021) work. Specifically, we use the pretrained DDPM models (Ho et al., 2020) on LSUN Bedroom and Church datasets to compare DDIM vs. DDIM-GMM-ORTHO-VUB samplers for different number of sampling steps (10, 20, 50 and 100).

Steps	10	20	50	100
DDIM	16.93	8.77	6.68	6.76
DDIM-GMM	**16.86**	**8.76**	**6.62**	**6.67**

Table 3: LSUN Bedroom. Comparison between DDIM and DDIM-GMM on the FID(↓) metric. $\eta = 0$ for both samplers.

Steps	10	20	50	100
DDIM	19.39	**12.33**	11.04	10.85
DDIM-GMM	**19.33**	12.37	**10.85**	**10.81**

Table 4: LSUN Church. Comparison between DDIM and DDIM-GMM on the FID(↓) metric. $\eta = 0$ for both samplers.

A.13 FID AND IS METRICS

In this section we list the metrics plotted in Section 4 in a tabular format. The numbers in bold emphasize improvement of the corresponding sampling method's metric if the difference in metric (FID or IS) is at least 1 unit from the worst result within the same group (same η and number of sampling steps). The number in parentheses denotes the scale parameter s that resulted in the best metric for the particular DDIM-GMM-* sampling method under a given setting. We omit the best s for ImageNet results.

A.13.1 CELEBAHQ

	Steps	10	20	50	100	1000
η						
0	DDIM	32.95	18.58	12.65	11.42	
0	DDIM-GMM-RAND	**28.01 (10)**	18.74 (1)	12.33 (1)	11.35 (1)	
0	DDIM-GMM-ORTHO	**27.97 (10)**	18.71 (1)	12.41 (1)	11.44 (1)	
0	DDIM-GMM-ORTHO-VUB	**27.94 (10)**	18.71 (1)	12.41 (1)	11.44 (1)	
0	DDIM-GMM-ORTHO-VUB*	**27.84 (10)**	18.53 (1)	12.62 (1)	11.42 (0.1)	
0.2	DDIM	33.74	19.48	12.79	11.41	
0.2	DDIM-GMM-RAND	**32.32 (10)**	**17.26 (10)**	12.80 (0.01)	11.36 (0.1)	
0.2	DDIM-GMM-ORTHO	**32.42 (10)**	**17.33 (10)**	12.79 (1)	11.37 (0.01)	
0.2	DDIM-GMM-ORTHO-VUB	**28.32 (10)**	19.18 (1)	12.58 (1)	11.29 (1)	
0.2	DDIM-GMM-ORTHO-VUB*	**27.68 (10)**	19.77 (0.01)	12.81 (1)	11.34 (1)	
0.5	DDIM	39.04	22.01	13.99	12.05	
0.5	DDIM-GMM-RAND	**37.15 (10)**	**20.66 (10)**	**12.95 (10)**	12.26 (1)	
0.5	DDIM-GMM-ORTHO	**37.27 (10)**	**20.78 (10)**	**12.98 (10)**	12.25 (0.1)	
0.5	DDIM-GMM-ORTHO-VUB	**31.00 (10)**	**20.65 (10)**	13.87 (1)	12.00 (1)	
0.5	DDIM-GMM-ORTHO-VUB*	**31.42 (10)**	**20.89 (10)**	14.09 (1)	11.91 (1)	
1.0	DDIM	68.67	39.20	21.53	16.09	
1.0	DDIM-GMM-RAND	**66.94 (10)**	**37.63 (10)**	**19.33 (10)**	**14.14 (10)**	
1.0	DDIM-GMM-ORTHO	**67.15 (10)**	**37.79(1)**	**19.36 (10)**	**14.37 (10)**	
1.0	DDIM-GMM-ORTHO-VUB	**60.65 (10)**	**28.03 (10)**	**15.97 (10)**	15.60 (1)	
1.0	DDIM-GMM-ORTHO-VUB*	**61.15 (10)**	**27.41 (10)**	**16.48 (10)**	15.94 (1)	
1.0	DDPM					11.59

Table 5: CelebAHQ (FID↓)

A.13.2 FFHQ

	Steps	10	20	50	100	1000
η						
0	DDIM	28.73	15.68	11.67	11.17	
0	DDIM-GMM-RAND	**26.55 (10)**	15.64 (1)	11.83 (0.01)	11.12 (0.01)	
0	DDIM-GMM-ORTHO	**26.46 (10)**	15.67 (1)	11.83 (0.01)	11.12 (0.01)	
0	DDIM-GMM-ORTHO-VUB	**26.46 (10)**	15.67 (1)	11.83 (0.01)	11.12 (0.01)	
0	DDIM-GMM-ORTHO-VUB*	**27.18 (10)**	15.43 (1)	11.77 (0.1)	11.24 (0.1)	
0.2	DDIM	29.33	15.83	11.66	11.07	
0.2	DDIM-GMM-RAND	29.27 (0.1)	16.03 (0.1)	11.87 (0.01)	10.92 (0.01)	
0.2	DDIM-GMM-ORTHO	29.09 (10)	16.01 (1)	11.87 (0.01)	10.92 (0.01)	
0.2	DDIM-GMM-ORTHO-VUB	**26.90 (10)**	15.96 (1)	11.87 (0.01)	10.91 (0.1)	
0.2	DDIM-GMM-ORTHO-VUB*	**27.35 (10)**	15.85 (0.01)	11.62 (0.01)	11.11 (0.01)	
0.5	DDIM	35.53	17.83	11.89	10.53	
0.5	DDIM-GMM-RAND	35.16 (10)	18.01 (10)	11.85 (1)	10.49 (0.01)	
0.5	DDIM-GMM-ORTHO	35.00 (10)	17.92 (10)	11.87 (0.1)	10.47 (0.1)	
0.5	DDIM-GMM-ORTHO-VUB	**29.03 (10)**	18.16 (1)	11.78 (1)	10.45 (0.1)	
0.5	DDIM-GMM-ORTHO-VUB*	**28.73 (10)**	17.89 (1)	11.92 (1)	10.81 (1)	
1.0	DDIM	81.88	37.09	15.45	11.33	
1.0	DDIM-GMM-RAND	**79.93**	**35.85 (10)**	15.17 (10)	11.50 (1)	
1.0	DDIM-GMM-ORTHO	**80.18 (10)**	**35.93 (10)**	15.03 (10)	11.51 (1)	
1.0	DDIM-GMM-ORTHO-VUB	**69.25 (10)**	**23.88 (1)**	15.44 (0.1)	11.44 (1)	
1.0	DDIM-GMM-ORTHO-VUB*	**67.72 (10)**	**23.76 (10)**	15.56 (1)	11.38 (0.1)	
1.0	DDPM					9.69

Table 6: FFHQ (FID↓)

A.13.3 IMAGENET

The FID and IS results for class-conditional models without classifier guidance are reported in Tables 7 and 8 respectively. The corresponding results with classifier and classifier-free guidance are in Tables 9-10 and Tables 11-12 respectively.

	Steps	10	20	50	100	1000
η						
0	DDIM	36.60	23.06	20.07	20.02	
0	DDIM-GMM-RAND	36.27	23.15	19.98	19.85	
0	DDIM-GMM-ORTHO	36.31	23.12	19.98	19.94	
0	DDIM-GMM-ORTHO-VUB	36.35	23.12	19.98	19.81	
0	DDIM-GMM-ORTHO-VUB*	36.27	23.08	20.04	19.70	
0.2	DDIM	37.18	23.18	20.07	19.39	
0.2	DDIM-GMM-RAND	37.04	23.17	20.13	19.50	
0.2	DDIM-GMM-ORTHO	37.14	23.15	20.15	19.75	
0.2	DDIM-GMM-ORTHO-VUB	36.93	23.16	20.19	19.62	
0.2	DDIM-GMM-ORTHO-VUB*	37.02	23.20	19.81	19.65	
0.5	DDIM	41.52	24.15	19.37	18.70	
0.5	DDIM-GMM-RAND	41.24	23.88	19.22	18.34	
0.5	DDIM-GMM-ORTHO	41.42	23.90	19.25	18.85	
0.5	DDIM-GMM-ORTHO-VUB	41.06	23.89	19.25	18.80	
0.5	DDIM-GMM-ORTHO-VUB*	41.46	23.81	19.27	18.69	
1.0	DDIM	71.54	32.26	18.69	16.95	
1.0	DDIM-GMM-RAND	70.91	32.25	18.71	16.81	
1.0	DDIM-GMM-ORTHO	71.52	32.22	18.73	16.91	
1.0	DDIM-GMM-ORTHO-VUB	**62.71**	32.10	18.64	16.75	
1.0	DDIM-GMM-ORTHO-VUB*	**63.02**	32.03	18.32	16.88	
1.0	DDPM					17.07

Table 7: Class-conditional ImageNet (FID↓)

	Steps	10	20	50	100	1000
η						
0	DDIM	37.66	51.70	57.60	57.38	
0	DDIM-GMM-RAND	37.79	51.89	57.13	57.98	
0	DDIM-GMM-ORTHO	37.86	51.88	57.03	57.12	
0	DDIM-GMM-ORTHO-VUB	37.85	51.88	57.03	57.52	
0	DDIM-GMM-ORTHO-VUB*	38.16	52.00	56.99	57.90	
0.2	DDIM	37.10	52.00	58.08	58.46	
0.2	DDIM-GMM-RAND	38.14	51.75	57.25	58.80	
0.2	DDIM-GMM-ORTHO	37.94	51.78	57.43	58.10	
0.2	DDIM-GMM-ORTHO-VUB	37.90	51.82	57.34	58.52	
0.2	DDIM-GMM-ORTHO-VUB*	37.38	52.47	58.34	58.04	
0.5	DDIM	34.52	51.15	58.71	61.03	
0.5	DDIM-GMM-RAND	34.46	51.10	59.32	60.91	
0.5	DDIM-GMM-ORTHO	34.52	50.84	59.06	59.65	
0.5	DDIM-GMM-ORTHO-VUB	**35.54**	51.14	59.08	60.70	
0.5	DDIM-GMM-ORTHO-VUB*	34.50	51.53	58.98	60.97	
1.0	DDIM	17.52	42.10	61.83	64.55	
1.0	DDIM-GMM-RAND	**18.52**	**43.13**	61.37	65.54	
1.0	DDIM-GMM-ORTHO	**18.58**	43.05	61.02	**65.82**	
1.0	DDIM-GMM-ORTHO-VUB	**24.25**	**44.01**	61.45	**65.97**	
1.0	DDIM-GMM-ORTHO-VUB*	**24.04**	42.69	61.99	**65.87**	
1.0	DDPM					66.18

Table 8: Class-conditional ImageNet (IS↑)

Steps		10		100		1000	
Guidance Scale		1	10	1	10	1	10
η							
0	DDIM	21.94	15.65	10.78	11.66		
0	DDIM-GMM-RAND	21.94	**11.32**	11.14	11.28		
0	DDIM-GMM-ORTHO	22.00	**11.26**	11.22	11.28		
0	DDIM-GMM-ORTHO-VUB	22.00	**11.26**	11.24	11.28		
0	DDIM-GMM-ORTHO-VUB*	21.95	**11.54**	10.81	11.31		
0.2	DDIM	22.39	15.83	10.59	11.55		
0.2	DDIM-GMM-RAND	22.25	**13.10**	10.50	11.58		
0.2	DDIM-GMM-ORTHO	22.11	**12.95**	10.50	11.57		
0.2	DDIM-GMM-ORTHO-VUB	22.43	**11.29**	10.79	11.20		
0.2	DDIM-GMM-ORTHO-VUB*	22.39	**11.50**	10.50	11.24		
0.5	DDIM	25.31	16.72	10.04	11.39		
0.5	DDIM-GMM-RAND	25.39	**15.36**	9.93	11.25		
0.5	DDIM-GMM-ORTHO	25.35	**15.27**	9.88	11.25		
0.5	DDIM-GMM-ORTHO-VUB	24.72	**11.79**	9.94	11.01		
0.5	DDIM-GMM-ORTHO-VUB*	24.96	**11.93**	9.91	10.96		
1.0	DDIM	47.61	26.09	8.97	11.48		
1.0	DDIM-GMM-RAND	46.55	**23.71**	8.94	**10.31**		
1.0	DDIM-GMM-ORTHO	46.45	**23.57**	8.91	**10.28**		
1.0	DDIM-GMM-ORTHO-VUB	**37.97**	**18.60**	8.82	11.38		
1.0	DDIM-GMM-ORTHO-VUB*	**38.08**	**18.68**	8.80	11.40		
1.0	DDPM					8.50	10.75

Table 9: Class-conditional ImageNet with classifier guidance(FID↓)

Steps		10		100		1000	
Guidance Scale		1	10	1	10	1	10
η							
0	DDIM	66.15	136.76	98.71	179.62		
0	DDIM-GMM-RAND	66.26	**161.19**	98.34	**181.43**		
0	DDIM-GMM-ORTHO	66.16	**161.47**	97.25	**181.33**		
0	DDIM-GMM-ORTHO-VUB	66.17	**161.45**	97.14	**181.33**		
0	DDIM-GMM-ORTHO-VUB*	66.48	**154.64**	98.75	180.60		
0.2	DDIM	66.17	135.01	99.4	179.86		
0.2	DDIM-GMM-RAND	69.51	**152.48**	99.60	**182.41**		
0.2	DDIM-GMM-ORTHO	**69.48**	**152.10**	100.12	**182.25**		
0.2	DDIM-GMM-ORTHO-VUB	65.35	**162.13**	99.00	**183.91**		
0.2	DDIM-GMM-ORTHO-VUB*	66.50	**156.31**	99.64	**182.60**		
0.5	DDIM	62.42	131.49	105.02	190.45		
0.5	DDIM-GMM-RAND	62.25	**138.43**	106.46	191.35		
0.5	DDIM-GMM-ORTHO	62.53	**139.64**	**106.18**	190.87		
0.5	DDIM-GMM-ORTHO-VUB	**67.45**	**161.57**	**106.74**	**192.72**		
0.5	DDIM-GMM-ORTHO-VUB*	**65.35**	**156.91**	**106.46**	**192.17**		
1.0	DDIM	35.19	86.76	116.59	207.78		
1.0	DDIM-GMM-RAND	36.78	**95.47**	116.54	207.52		
1.0	DDIM-GMM-ORTHO	**37.02**	**96.90**	**117.09**	208.34		
1.0	DDIM-GMM-ORTHO-VUB	**48.74**	**120.63**	**117.19**	**209.36**		
1.0	DDIM-GMM-ORTHO-VUB*	35.03	**118.57**	117.52	208.74		
1.0	DDPM					118.28	210.53

Table 10: Class-conditional ImageNet with classifier guidance(IS↑)

Steps		10		100		1000	
Guidance Scale		2.5	5	2.5	5	2.5	5
η							
0	DDIM	10.15	18.56	9.57	19.94		
0	DDIM-GMM-RAND	**6.90**	**16.77**	9.23	19.83		
0	DDIM-GMM-ORTHO	**6.90**	**16.64**	9.20	19.84		
0	DDIM-GMM-ORTHO-VUB	**6.94**	**16.61**	9.13	19.66		
0	DDIM-GMM-ORTHO-VUB*	**6.72**	**16.14**	9.22	19.77		
0.2	DDIM	10.35	18.60	9.67	20.29		
0.2	DDIM-GMM-RAND	**8.52**	17.90	9.70	20.17		
0.2	DDIM-GMM-ORTHO	**8.54**	17.85	9.73	20.18		
0.2	DDIM-GMM-ORTHO-VUB	**7.06**	**16.78**	9.27	20.04		
0.2	DDIM-GMM-ORTHO-VUB*	**6.75**	**16.49**	9.38	20.05		
0.5	DDIM	11.15	19.13	10.70	21.41		
0.5	DDIM-GMM-RAND	10.28	18.61	10.60	**15.73**		
0.5	DDIM-GMM-ORTHO	10.29	18.54	10.56	**15.73**		
0.5	DDIM-GMM-ORTHO-VUB	**7.59**	**17.91**	10.27	21.10		
0.5	DDIM-GMM-ORTHO-VUB*	**7.44**	**17.51**	10.38	21.17		
1.0	DDIM	17.50	20.42	12.97	23.56		
1.0	DDIM-GMM-RAND	**15.95**	19.91	**10.84**	**22.38**		
1.0	DDIM-GMM-ORTHO	**15.94**	19.92	**10.86**	**22.26**		
1.0	DDIM-GMM-ORTHO-VUB	**12.38**	19.70	12.80	22.94		
1.0	DDIM-GMM-ORTHO-VUB*	**12.34**	19.64	12.80	23.59		
1.0	DDPM					12.61	23.38

Table 11: Class-conditional ImageNet with classifier free guidance(FID↓)

Steps		10		100		1000	
Guidance Scale		2.5	5	2.5	5	2.5	5
η							
0	DDIM	196.73	296.89	237.98	321.59		
0	DDIM-GMM-RAND	**209.90**	**318.73**	238.63	321.68		
0	DDIM-GMM-ORTHO	**207.65**	**320.66**	238.97	**322.90**		
0	DDIM-GMM-ORTHO-VUB	**207.85**	**319.13**	**240.88**	**323.84**		
0	DDIM-GMM-ORTHO-VUB*	**200.89**	**316.90**	238.97	321.40		
0.2	DDIM	196.15	298.19	241.48	323.87		
0.2	DDIM-GMM-RAND	**209.22**	**311.00**	239.52	323.23		
0.2	DDIM-GMM-ORTHO	**209.02**	**313.10**	238.85	322.96		
0.2	DDIM-GMM-ORTHO-VUB	**208.91**	**318.41**	**243.14**	**326.02**		
0.2	DDIM-GMM-ORTHO-VUB*	**204.41**	**318.19**	241.09	**324.86**		
0.5	DDIM	193.44	297.65	250.97	330.86		
0.5	DDIM-GMM-RAND	**198.14**	**306.79**	251.02	331.14		
0.5	DDIM-GMM-ORTHO	**197.84**	**305.75**	250.27	331.16		
0.5	DDIM-GMM-ORTHO-VUB	**211.93**	**319.87**	**251.72**	**333.34**		
0.5	DDIM-GMM-ORTHO-VUB*	**210.79**	**320.60**	**251.74**	**333.29**		
1.0	DDIM	148.96	281.86	272.39	345.19		
1.0	DDIM-GMM-RAND	**157.54**	**294.47**	270.73	**354.69**		
1.0	DDIM-GMM-ORTHO	**158.41**	**294.03**	270.28	**355.00**		
1.0	DDIM-GMM-ORTHO-VUB	**187.49**	**309.10**	**273.73**	**349.01**		
1.0	DDIM-GMM-ORTHO-VUB*	**185.85**	**310.05**	271.89	**346.74**		
1.0	DDPM					277.40	349.48

Table 12: Class-conditional ImageNet with classifier free guidance(IS↑)

A.14 QUALITATIVE RESULTS

In this section we show some qualitative results of sampling with the proposed approach compared to original DDIM. Specifically, we use the DDIM-GMM-ORTHO-VUB* (Section 3.1.3) method to obtain the samples and refer to them with the label DDIM-GMM for brevity. In Fig. 12 we plot samples from the class conditional model trained on ImageNet with 10 sampling steps for both DDIM (left) and DDIM-GMM (right). The top and bottom group of images correspond to input class labels "pelican" and "cairn terrier" respectively. With only a few sampling steps, the semantic concept seems to emerge clearer in the images obtained from DDIM-GMM sampler than in the ones from DDIM sampler. Under this setting, the IS metric for the DDIM-GMM sampler is significantly higher than that of DDIM (see Table 8, Appendix A.13.3). See Appendix A.15 for more comparisons.

Figure 12: **Class-conditional ImageNet, 10 sampling steps.** Random samples from the class-conditional ImageNet model using DDIM (left) and DDIM-GMM (right) sampler conditioned on the class labels *pelican* (top) and *cairn terrier* (bottom) respectively. 10 sampling steps are used for each sampler ($\eta = 1$).

A.15 MORE QUALITATIVE RESULTS

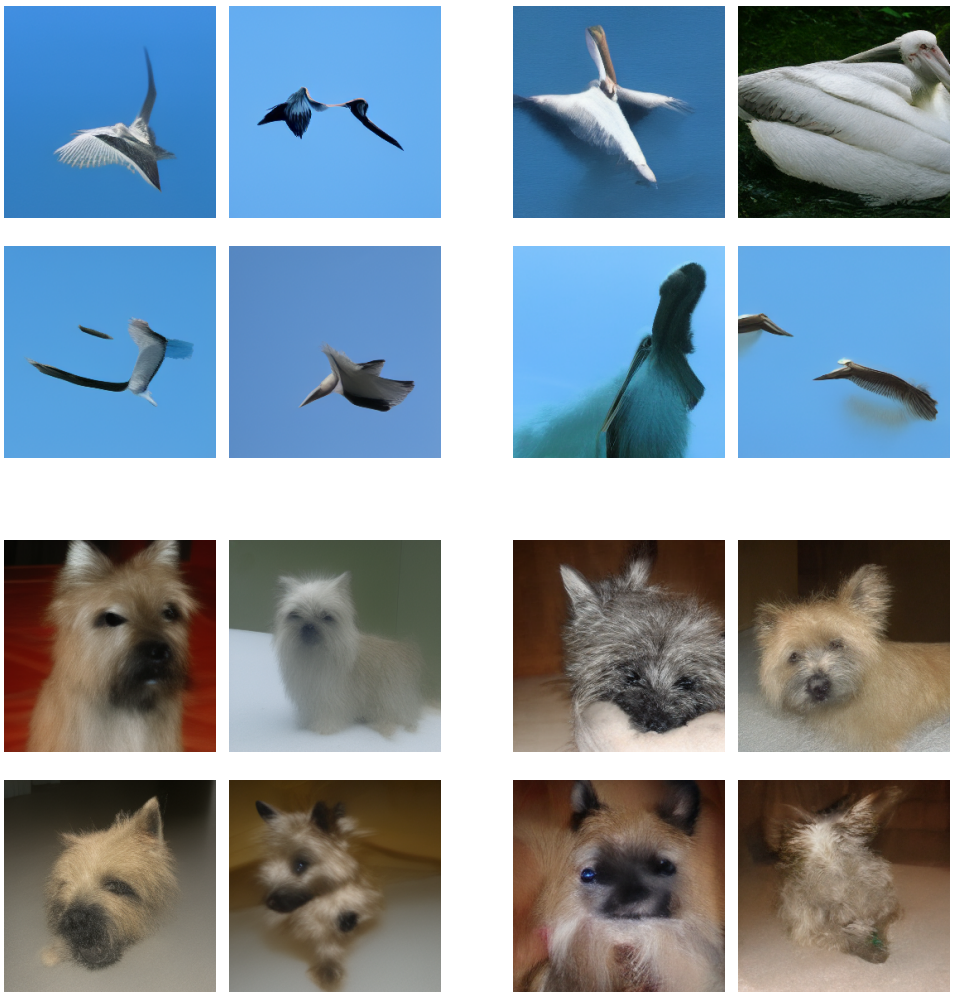

Figure 13: **Class-conditional ImageNet with classifier guidance, 10 sampling steps.** Random samples from the class-conditional ImageNet model using DDIM (left) and DDIM-GMM (right) sampler conditioned on the class labels *pelican* (top) and *cairn terrier* (bottom) respectively. 10 sampling steps are used for each sampler with a classifier guidance weight of 10 ($\eta = 1$).

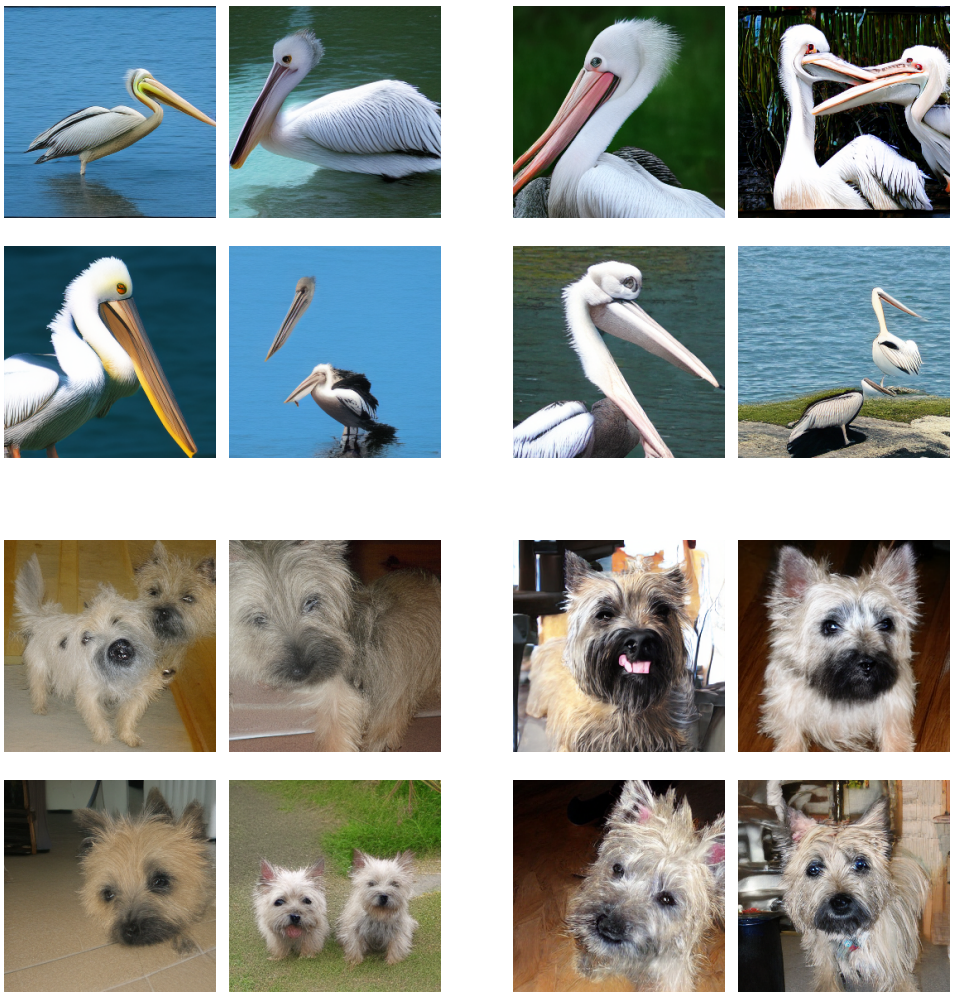

Figure 14: **Class-conditional ImageNet with classifier-free guidance, 10 sampling steps.** Random samples from the class-conditional ImageNet model using DDIM (left) and DDIM-GMM (right) sampler conditioned on the class labels *pelican* (top) and *cairn terrier* (bottom) respectively. 10 sampling steps are used for each sampler with a classifier free guidance weight of 5 ($\eta = 0$).