Unleashing Transformers: Parallel Token Prediction with Discrete Absorbing Diffusion for Fast High-Resolution Image Generation from Vector-Quantized Codes

Sam Bond-Taylor[1⋆], Peter Hessey[1⋆], Hiroshi Sasaki[1],
Toby P. Breckon[1,2], and Chris G. Willcocks[1]

[1] Department of Computer Science, Durham University, Durham, UK
[2] Department of Engineering, Durham University, Durham, UK
{samuel.e.bond-taylor, peter.hessey, hiroshi.sasaki,
toby.breckon, christopher.g.willcocks}@durham.ac.uk

Abstract. Whilst diffusion probabilistic models can generate high quality image content, key limitations remain in terms of both generating high-resolution imagery and their associated high computational requirements. Recent Vector-Quantized image models have overcome this limitation of image resolution but are prohibitively slow and unidirectional as they generate tokens via element-wise autoregressive sampling from the prior. By contrast, in this paper we propose a novel discrete diffusion probabilistic model prior which enables parallel prediction of Vector-Quantized tokens by using an unconstrained Transformer architecture as the backbone. During training, tokens are randomly masked in an order-agnostic manner and the Transformer learns to predict the original tokens. This parallelism of Vector-Quantized token prediction in turn facilitates unconditional generation of globally consistent high-resolution and diverse imagery at a fraction of the computational expense. In this manner, we can generate image resolutions exceeding that of the original training set samples whilst additionally provisioning per-image likelihood estimates (in a departure from generative adversarial approaches). Our approach achieves state-of-the-art results in terms of the manifold overlap metrics Coverage (LSUN Bedroom: 0.83; LSUN Churches: 0.73; FFHQ: 0.80) and Density (LSUN Bedroom: 1.51; LSUN Churches: 1.12; FFHQ: 1.20), and performs competitively on FID (LSUN Bedroom: 3.64; LSUN Churches: 4.07; FFHQ: 6.11) whilst offering advantages in terms of both computation and reduced training set requirements.

Keywords: generative model, diffusion, high-resolution image synthesis

1 Introduction

Artificially generating plausible photo-realistic images, at ever higher resolutions, has long been a goal when designing deep generative models. Recent advance-

[⋆] Authors contributed equally. Source code for this work is available at https://github.com/samb-t/unleashing-transformers.

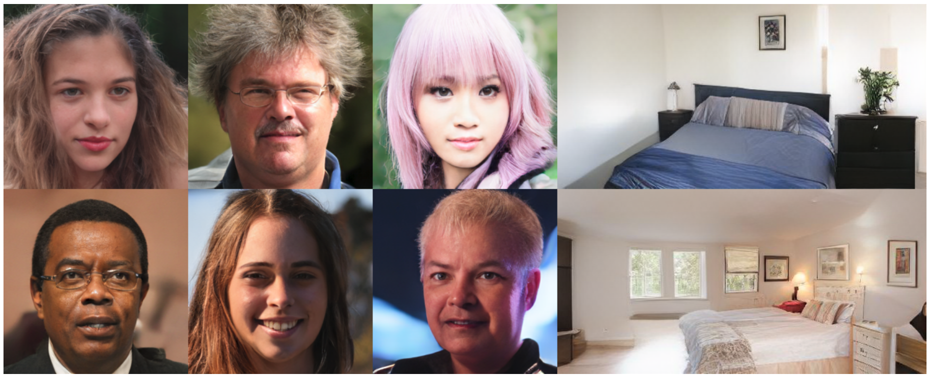


Fig. 1: Our approach uses a discrete diffusion to quickly generate high quality images optionally larger than the training data (right).

ments have benefited fields such as medical image synthesis [23], computer graphics [11,91], image editing [52], image translation [77], and super-resolution [33].

These methods can be divided into five main classes [5], each making different trade-offs to scale to high resolutions. Techniques to scale Generative Adversarial Networks (GANs) [25] include progressive growing [42], large batches [8], and regularisation [53,56]. Variational Autoencoders (VAEs) [49] can be scaled by building complex priors [12,63,84] and correcting samples [89]. Autoregressive approaches can make independence assumptions [69] or partition dimensions [55]. Normalizing Flows use multi-scale architectures [50], while diffusion models can be scaled using SDEs [81] and cascades [33]. Each of these have their own drawbacks, such as unstable training, slow sampling, and lack of global context.

Of particular interest to this work is the popular Transformer architecture [86] which models long distance relationships using a powerful parallelisable attention mechanism. By constraining the Transformer architecture to attend a fixed unidirectional ordering of tokens, they can be used to parameterise a generative autoregressive model [13,64]. However, image data does not conform to such a structure and hence this bias limits the representation ability and unnecessarily restricts the sampling process to be sequential and slow.

Addressing these issues, our main contributions are:

- We propose a parallel token prediction approach for generating Vector-Quantized images allowing much faster sampling than autoregressive models.
- Our approach is able to generate globally consistent images at resolutions exceeding that of the original training data by aggregating multiple context windows, allowing for much larger context regions.
- Our approach demonstrates state-of-the art performance on three benchmark datasets in terms of Density (LSUN Bedroom: 1.51; LSUN Churches: 1.12; FFHQ: 1.20) and Coverage (Bedroom: 0.83; Churches: 0.73; FFHQ: 0.80), and is competitive on FID (Bedroom: 3.64; Churches: 4.07; FFHQ: 6.11).

2 Prior Work

Extensive work in deep generative modelling [5] and self-supervised learning [20] laid the foundations for this research, which we review here in terms of both existing models (Sections 2.1-2.4) and Transformer architectures (Section 2.5).

2.1 Autoregressive Models

Autoregressive models are a family of powerful generative models capable of directly maximising the likelihood of the data on which they are trained. These models have achieved impressive image generation results, however, their sequential nature limits them to relatively low dimensional data [14, 41, 62, 71, 76, 85].

The training and inference process for autoregressive models is based on the chain rule. By decomposing inputs into components $\boldsymbol{x} = \{x_1, ..., x_n\}$, an autoregressive model with parameters θ can generate new latent samples sequentially:

$$p_\theta(\boldsymbol{x}) = p_\theta(x_1, x_2, ..., x_n) = \prod\nolimits_{i=1}^{n} p_\theta(x_i|x_1, ..., x_{i-1}). \tag{1}$$

For many tasks, appropriate input orderings are not obvious; since the receptive field is limited to previous tokens, this can significantly affect sample quality.

2.2 Vector-Quantized Image Models

To scale autoregressive models to high-resolution data, Vector-Quantized image models can be used. These learn a highly compressed discrete representation taking advantage of an information rich codebook [63]. A convolutional encoder downsamples images $\boldsymbol{x}$ to a smaller spatial resolution, $E(\boldsymbol{x}) = \{\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_L\} \in \mathbb{R}^{L \times D}$. A simple quantisation approach is to use the argmax operation which maps continuous encodings to their closest elements in a finite codebook of vectors [63]. Specifically, for a codebook $\mathcal{C} \in \mathbb{R}^{K \times D}$, where K is the number of discrete codes in the codebook and D is the dimension of each code, each $\boldsymbol{e}_i$ is mapped via a nearest-neighbour lookup onto a discrete codebook value, $\boldsymbol{c}_j \in \mathcal{C}$:

$$\boldsymbol{z}_q = \{\boldsymbol{q}_1, \boldsymbol{q}_2, ..., \boldsymbol{q}_L\} \text{ , where } \boldsymbol{q}_i = \min_{\boldsymbol{c}_j \in \mathcal{C}} \|\boldsymbol{e}_i - \boldsymbol{c}_j\|. \tag{2}$$

As this operation is non-differentiable, the straight-through gradient estimator [3] is used to approximate gradients resulting in bias. The quantized latents are fed through a decoder $\hat{\boldsymbol{x}} = G(\boldsymbol{z}_q)$ to reconstruct the input based on a perceptual reconstruction loss [22, 92]; this process is trained by minimising the loss $\mathcal{L}_{\text{VQ}}$,

$$\mathcal{L}_{\text{VQ}} = \mathcal{L}_{\text{rec}} + \|\text{sg}[E(\boldsymbol{x})] - \boldsymbol{z}_q\|_2^2 + \beta \|\text{sg}[\boldsymbol{z}_q] - E(\boldsymbol{x})\|_2^2. \tag{3}$$

2.3 Discrete Energy-Based Models

Since the causal nature of autoregressive models limits their representation ability, other approaches with less constrained architectures have begun to outperform them even on likelihood [48]. Energy-based models (EBMs) are an enticing method for representing discrete data as they permit unconstrained architectures with global context. Implicit EBMs define an unnormalised distribution over data that is typically learned through contrastive divergence [19, 31]. Unfortunately, sampling EBMs using Gibbs sampling is impractical for high dimensional discrete data. However, incorporating gradients can reduce mixing times [27].

Similar to autoregressive models, masked language models (MLMs) such as BERT [15] model the conditional probability of the data. However, these are trained bidirectionally by randomly masking a subset of tokens from the input sequence, allowing a much richer context than autoregressive approaches. Some attempts have been made to define an implicit energy function using the conditional probabilities [87], however, obtaining true samples leads to very long sample times and we found them to be ineffective at modelling longer sequences [26].

2.4 Discrete Denoising Diffusion Models

Diffusion models [32, 80] define a Markov chain $q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ that gradually destroys data $\boldsymbol{x}_0$ by adding noise over a fixed number of steps T so that $\boldsymbol{x}_T$ contains little to no information about $\boldsymbol{x}_0$ and can be easily sampled. The reverse procedure is a generative model that gradually denoises towards the data distribution $p_\theta(\boldsymbol{x}_{0:T}) = p_\theta(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$, learned by optimising the variational upper bound on negative log-likelihood, with t^{th} term

$$\mathbb{E}_{q(\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)} \left[D_{KL}(q(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}, \boldsymbol{x}_0) \parallel p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})) \right], \tag{4}$$

where sampling from the reverse process is not required during training. In continuous spaces, distributions are typically parameterised as Normal distributions.

Discrete diffusion models [1, 36, 80] constrain the state space so that $\boldsymbol{x}_t$ is a discrete random variable falling into one of K categories. As such, the forward process can be represented as categorical distributions $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \text{Cat}(\boldsymbol{x}_t; \boldsymbol{p} = \boldsymbol{x}_{t-1}\boldsymbol{Q}_t)$ for one-hot $\boldsymbol{x}_{t-1}$ where $\boldsymbol{Q}_t$ is a matrix denoting the probabilities of moving to each successive state. $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ can be expressed as $q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \text{Cat}(\boldsymbol{x}_t; \boldsymbol{p} = \overline{\boldsymbol{Q}}_t)$ where $\overline{\boldsymbol{Q}}_t = \boldsymbol{x}_0\boldsymbol{Q}_1\boldsymbol{Q}_2 \cdots \boldsymbol{Q}_t$, therefore scaling is simple if $\overline{\boldsymbol{Q}}_t$ can be expressed in closed form. Transition processes include moving states with some low uniform probability [36], moving to nearby states with some probability based on similarity or distance, and of particular interest to this work, masking inputs similar to generative MLMs.

2.5 Transformers

Transformers [86] have made a huge impact across many fields [30] due to their power and flexibility. They are based on self-attention, a function which allows interactions with strong gradients between all inputs, irrespective of their spatial relationships. This procedure (Eqn. 5) encodes inputs as key-value pairs, where values $\boldsymbol{V}$ represent embedded inputs and keys $\boldsymbol{K}$ act as an indexing method, subsequently, a set of queries $\boldsymbol{Q}$ are used to select which values to observe:

$$\text{Attn}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V}. \tag{5}$$

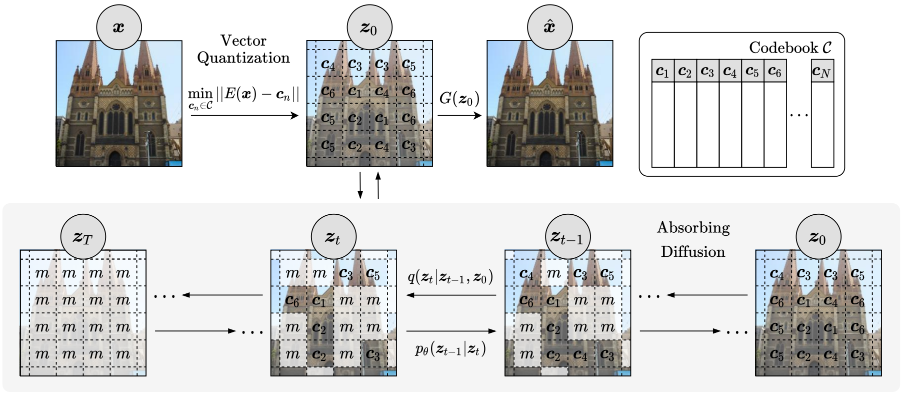


Fig. 2: Our approach uses a discrete absorbing diffusion model to represent Vector-Quantized images allowing fast high-resolution image generation. Specifically, after compressing images to an information-rich discrete space, elements are randomly masked and an unconstrained Transformer is trained to denoise the data, using global context to ensure samples are consistent and high quality.

While this allows long distance dependencies to be learned, complexity increases with sequence length quadratically, making scaling difficult. Approaches to mitigate this include independence assumptions [69], sparsity [13], and low rank [46].

3 Method

Modelling Vector-Quantized image representations with autoregressive models has a number of downsides, namely the slow sequential nature of sampling and the requirement to choose an input ordering which ignores the 2D structure of images thereby restricting modelling ability. To address these problems, we propose using a discrete diffusion model to represent Vector-Quantized image representations; this is visualised in Fig. 2. We hypothesise that by removing the autoregressive constraint, allowing bidirectional context when generating samples, not only will it be possible to speed up sampling, but an improved feature representation will be learned, enabling higher quality image generation.

3.1 Sampling Globally Coherent Latents

Once the training data is encoded as discrete, integer-valued latents $\boldsymbol{z} \in \mathbb{Z}^D$, a discrete diffusion model can be used to learn the distribution over this highly compressed space. Specifically, we use the absorbing state diffusion [1] where in each forward time step t, each discrete latent at coordinate i, $[\boldsymbol{z}]_i$, is independently either kept the same or masked out entirely with probability $\frac{1}{t}$; the reverse process gradually unveils these masks. In this formulation, the transition matrix is defined as $\boldsymbol{Q}_t = (1-\beta_t)I + \beta_t \mathbb{1}e_m^T$ where e_m is a vector with a one on mask states m and zeros elsewhere, and the beta schedule is $\beta_t = \frac{1}{T-t+1}$. Rather than directly approximating $p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t)$, training stochasticity is reduced by predicting $p_\theta(\boldsymbol{z}_0|\boldsymbol{z}_t)$ [32]. In this case, the variational bound reduces to

$$\mathbb{E}_{q(\boldsymbol{z}_0)}\left[\sum_{t=1}^{T}\frac{1}{t}\,\mathbb{E}_{q(\boldsymbol{z}_t|\boldsymbol{z}_0)}\Big[\sum_{[\boldsymbol{z}_t]_i=m}\log p_\theta([\boldsymbol{z}_0]_i|\boldsymbol{z}_t)\Big]\right]. \tag{6}$$

With p_θ modelled using multinomial distributions, a temperature $\tau < 1$ can be applied to the logits to improve sample quality at the expense of diversity.

Unlike, uniform diffusion, absorbing diffusion is an effective strategy for Vector-Quantized image modelling as noisy elements are removed entirely rather than being changed to a different value which in the discrete case may be unrelated but are much less easy to identify. Gaussian and token distance transitions which change states based on embedding distances are similarly ineffective as Vector-Quantized latents are not ordinal meaning that state changes can significantly change tokens' semantics. This effectiveness is further evidenced by the success of BERT [15] which similarly learns to denoise randomly masked data.

Architecture Esser et al. [22] demonstrated that in the autoregressive case, Transformers [86] are better suited for modelling Vector-Quantized images than convolutional approaches due to the importance of long-distance relationships in this compressed form. As such, we use transformers to model the prior, but without the architectural restrictions imposed by autoregressive approaches.

Fast Sampling Because the diffusion model is trained to predict $p(\boldsymbol{z}_0|\boldsymbol{z}_t)$, it is possible to sample skipping an arbitrary number of time steps k, $p_\theta(\boldsymbol{z}_{t-k}|\boldsymbol{z}_t)$, allowing sampling in significantly fewer steps than autoregressive approaches.

3.2 Addressing Gradient Variance

When inputs are very noisy (at time steps close to T), denoising is difficult and the stochastic training results in gradients with high variance. As such, in practice continuous diffusion models are trained to estimate the noise rather than directly predict the denoised data, significantly reducing the variance. Unfortunately, no relevant reparameterisation currently exists for discrete distributions [36]. Instead, we address this problem by reweighting the ELBO based on the information available at time t, $\frac{T-t+1}{T}$ [1], so that components of the loss at time steps closer to T are weighted less than earlier steps. This effectively alters the learning rate based on gradient variance, improving convergence,

$$\mathbb{E}_{q(\boldsymbol{z}_0)}\left[\sum\nolimits_{t=1}^{T}\frac{T-t+1}{T}\,\mathbb{E}_{q(\boldsymbol{z}_t|\boldsymbol{z}_0)}\Big[\sum\nolimits_{[\boldsymbol{z}_t]_i=m}\log p_\theta([\boldsymbol{z}_0]_i|\boldsymbol{z}_t)\Big]\right]. \tag{7}$$

This is equivalent to the loss obtained by assuming the posterior does not have access to $\boldsymbol{z}_t$, i.e. if the t–1^{th} loss term is $D_{KL}(q(\boldsymbol{z}_{t-1}|\boldsymbol{z}_0) \parallel p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t))$ (proof in Appendix B). Since we predict $\boldsymbol{z}_0$ this assumption does not harm the training.

3.3 Generating High-Resolution Images

Using convolutions to build Vector-Quantized image models encourages latents to be highly spatially correlated with generated images. It is therefore possible to construct essentially arbitrarily sized images by generating latents with the required shape. We propose an approach that allows globally consistent images substantially larger than those in the training data to be generated.

First, a large a by b array of mask tokens, $\bar{\boldsymbol{z}}_T = m^{a\times b}$, is initialised that corresponds to the size of image we wish to generate. In order to capture the maximum context when approximating $\bar{\boldsymbol{z}}_0$ we apply the denoising network to all subsets of $\bar{\boldsymbol{z}}_t$ with the same spatial size as the usual inputs of the network, aggregating estimates at each location. Specifically, using $c_j(\bar{\boldsymbol{z}}_t)$ to represent local subsets, we approximate the denoising distribution as a mixture,

$$p([\bar{\boldsymbol{z}}_0]_i|\bar{\boldsymbol{z}}_t) \approx \frac{1}{Z}\sum_j p([\bar{\boldsymbol{z}}_0]_i|c_j(\bar{\boldsymbol{z}}_t)), \tag{8}$$

where the sum is over subsets c_j that contain the i^{th} latent and Z is the normalising constant. For extremely large images, this can require a very large number of function evaluations, however, the sum can be approximated by striding over latents with a step > 1 or by randomly selecting positions.

3.4 Improving Code Representations

There are various options to obtain high-quality image representations including using large numbers of latents and codes [67] or building a hierarchy of latent variables [68]. We use the adversarial framework proposed by Esser et al. [22] to achieve higher compression rates with high-quality codes using only a single GPU, without tying our approach to the characteristics typically associated with generative adversarial models. Additionally, we apply differentiable augmentations T, such as translations and colour jitter, to all discriminator inputs; this has proven to be effective at improving sample quality across methods [41,93]. The overall loss $\mathcal{L}$ is a linear combination of $\mathcal{L}_{\text{VQ}}$, the Vector-Quantized loss, and $\mathcal{L}_{\text{G}}$ which uses a discriminator D to assess realism based on an adaptive weight λ. On some datasets, λ can grow to extremely large values hindering training. We find simply clamping λ at a maximum value $\lambda_{\max} = 1$ an effective solution that stabilises training,

$$\mathcal{L} = \min_{E,G,\mathcal{C}} \max_D \mathbb{E}_{p_d}\left[\,\mathcal{L}_{\text{VQ}} + \lambda\,\mathcal{L}_{\text{G}}\,\right], \tag{9a}$$

$$\lambda = \min\left(\frac{\nabla_{G_L}[\mathcal{L}_{\text{rec}}]}{\nabla_{G_L}[\mathcal{L}_{\text{G}}] + \delta}, \lambda_{\max}\right), \tag{9b}$$

$$\mathcal{L}_{\text{G}} = \log D(T(\boldsymbol{x})) + \log(1 - D(T(\hat{\boldsymbol{x}}))). \tag{9c}$$

The argmax quantisation approach can result in codebook collapse, where some codes are never used; while other quantisation methods can reduce this [17,40,54,67], we found argmax to yield the highest reconstruction quality.

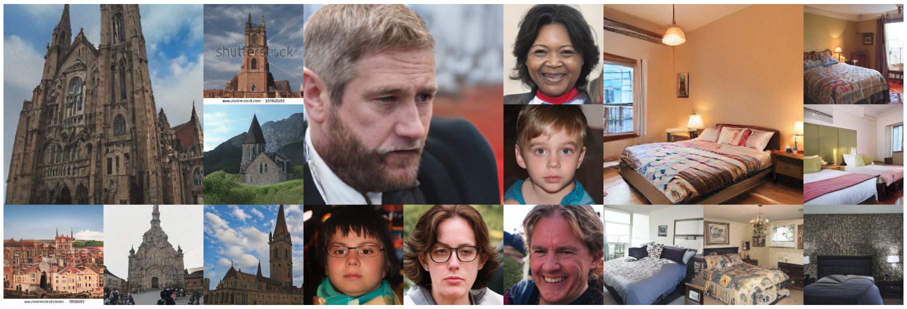


Fig. 3: Samples from our models trained on 256x256 datasets: LSUN Churches, FFHQ, and LSUN Bedroom.

4 Evaluation

We evaluate our approach on three high-resolution 256x256 datasets: LSUN Bedroom, LSUN Churches [90], and FFHQ [44]. Sec. 4.1 evaluates the quality of samples from our proposed model. Sec. 4.2 demonstrates the representation abilities of absorbing diffusion models applied to the learned discrete latent spaces, including how sampling can be sped up, improvements over equivalent autoregressive models, and the effect of our reweighted ELBO. Finally, Sec. 4.3 evaluates our Vector-Quantized image model.

In all experiments, our absorbing diffusion model parameterised with an 80M parameter Transformer Encoder [86] is applied to 16×16 latents discretised to a codebook with 1024 entries and optimised using the Adam optimiser [47]. While, as noted by Esser et al. [22], a GPT2-medium [66] architecture (307M parameters) fits onto a GPU with 12GB of VRAM, in practice this requires small batch sizes and learning rates making training in reasonable times impractical. More details can be found in Appendix A. Source code is available here.

4.1 Sample Quality

In this section we evaluate our model quantitatively and qualitatively. In contrast to other multi-step methods, our approach allows sampling in the fewest steps. Samples can be found in Figs. 3 and 5 which are high quality and diverse.

Limitations of the FID Metric FID is a popular choice for evaluating sample quality, it has been found to correlate well with image quality and is efficient to calculate. However, it unrealistically approximates the data distribution as Gaussian in embedding space and is insensitive to the global structure of the data distribution [83]. For likelihood models, calculating NLL is possible instead; by fine tuning our approach to model pixels as Gaussians, likelihood can be estimated as $p(\boldsymbol{x}) \geq p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$ [63], giving 2.72BPD on 5-bit FFHQ. However, likelihood does not correlate well with quality [82]. Other approaches that address these issues [6] include PPL [44], which assesses sample consistency through

	LSUN Churches				LSUN Bedroom				FFHQ			
Model	P ↑	R ↑	D ↑	C ↑	P ↑	R ↑	D ↑	C ↑	P ↑	R ↑	D ↑	C ↑
DCT [58]	0.60	0.48	-	-	0.44	**0.56**	-	-	0.51	0.40	-	-
TT [22]	0.67	0.29	1.08	0.60	0.61	0.33	1.15	0.75	0.64	0.29	0.89	0.5
VDVAE [12]	-	-	-	-	-	-	-	-	0.59	0.20	0.80	0.50
PGGAN [42]	0.61	0.38	0.83	0.63	0.43	0.40	0.70	0.64	-	-	-	-
StyleGAN [44]	-	-	-	-	0.55	0.48	0.96	0.80	-	-	-	-
StyleGAN2 [45]	0.60	0.43	0.83	0.68	-	-	-	-	0.69	0.40	1.12	0.80
ProjGAN [78]	0.56	**0.53**	0.65	0.64	0.55	0.46	0.90	0.79	0.66	0.46	0.98	0.77
Ours ($\tau = 1.0$)	0.70	0.42	**1.12**	0.73	0.64	0.38	1.27	0.81	0.69	0.48	1.06	0.77
Ours ($\tau = 0.9$)	**0.71**	0.45	1.07	**0.74**	**0.67**	0.38	**1.51**	**0.83**	**0.73**	**0.48**	**1.20**	**0.80**

Table 1: Precision (P), Recall (R), Density (D), and Coverage (C) [51,57,75] for approaches trained on LSUN Churches, LSUN Bedroom, and FFHQ.

latent interpolations; IMD [83], which uses all moments making it sensitive to global structure; and MTD [2], which compares image manifolds.

PRDC In this work, we evaluate using Precision (P) and Recall (R) [75] approaches (Tab. 1) which, unlike FID, evaluate sample quality and diversity separately by quantifying the overlap between the data and sample distributions, and have been used in similar recent work assessing high-resolution image generation [38,45,58,68]. Precision is the expected likelihood of fake samples lying on the data manifold and recall vice versa. These metrics are computed by approximating the data and sample manifolds as hyper-spheres around data and sample points respectively; manifold $\mathrm{m}(X_1, \dots, X_N) = \bigcup_{i=1}^{N} B(X_i, \mathrm{NND}_k(X_i))$, where $B(x, r)$ is a hypersphere around x with radius r and NND_k is k^{th} nearest neighbour distance [51]. While modelling manifolds as hyperspheres is a flawed assumption, it is beneficial to evaluate on multiple metrics to obtain a more accurate representation of performance. We also calculate Density (D) and Coverage (C) which are modifications to Precision and Recall respectively that address manifold overestimation [57]. Formally, these metrics can be defined as,

$$\mathrm{P} = \frac{1}{M} \sum_{j=1}^{M} 1_{Y_j \in \mathrm{m}(X_1, \dots, X_N)}, \tag{10a}$$

$$\mathrm{D} = \frac{1}{kM} \sum_{j=1}^{M} \sum_{i=1}^{N} 1_{Y_j \in B(X_i, \mathrm{NND}_k(X_i))}, \tag{10b}$$

$$\mathrm{R} = \frac{1}{N} \sum_{i=1}^{N} 1_{X_i \in \mathrm{m}(Y_1, \dots, Y_M)}, \tag{10c}$$

$$\mathrm{C} = \frac{1}{N} \sum_{i=1}^{N} 1_{\exists j \mathrm{s.t} Y_j \in B(X_i, \mathrm{NND}_k(X_i))}. \tag{10d}$$

Due to limited computing resources, we are unable to provide Density and Coverage scores for DCT [58] and PRDC scores for StyleGAN2 on LSUN Bedroom since training on a standard GPU would take more than 30 days, much more than the 10 days to train our models. On LSUN our approach achieves the highest Precision, Density, and Coverage; indicating that the data and sample manifolds

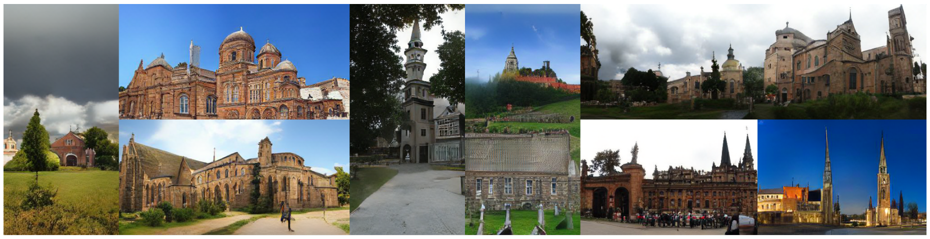


Fig. 4: Our method allows unconditional images larger than those seen during training to be generated by applying the denoising network to all subsets of the image, aggregating probabilities to encourage global continuity.

have the most overlap. On FFHQ our approach achieves the highest Precision and Recall. When sampling with lower temperatures to improve FID, generative models generally trade precision and recall [45, 68]; since we also calculate FID with $\tau = 0.9$, we evaluate the effect on PRDC. In almost all cases this improves scores, indicating that more samples in data regions, increasing overlap.

FID In Tab. 2 we calculate the Fréchet Inception Distance (FID) of samples from our models using torch-fidelity [61]. Using a fraction of the parameters of other Vector-Quantized image models, our approach achieves much lower FID.

Higher Resolution Fig. 4 shows samples generated at higher resolutions (up to 768×256) than the observed training data using the method described in Sec. 3.3 with $\tau = 0.8$. Even at larger scales we observe high-quality, diverse, and consistent samples.

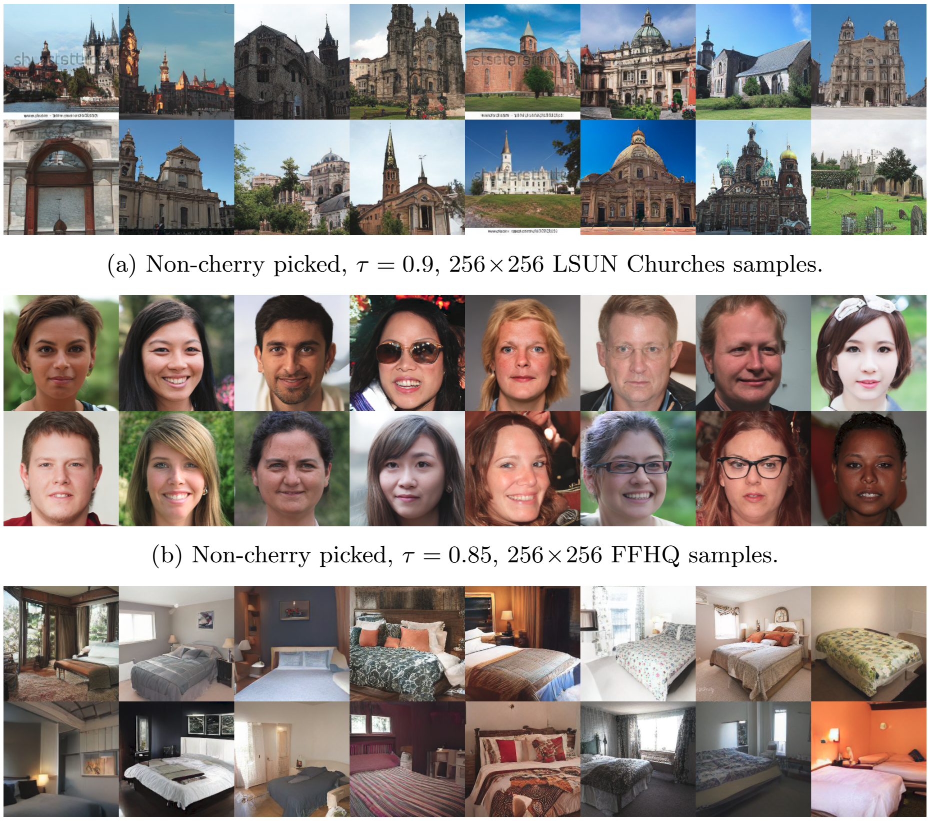
Method	Params	Bed	Church	FFHQ
DDPM [32]	114M	6.36	7.89	-
DCT [58]	448M	6.40	7.56	-
VDVAE [12]	115M	-	-	28.5
TT [21, 22]	600M	6.35	7.81	9.6
I-BART [21]	2.1B	5.51	7.32	9.57
PGGAN [42]	47M	8.34	6.42	-
SGAN2 [45]	60M	2.35	3.86	3.8
ADM [16]	552M	1.90	-	-
ProjGAN [78]	106M	1.52	1.59	3.39
Ours ($\tau = 1.0$)	145M	5.07	5.58	7.12
Ours ($\tau = 0.9$)	145M	3.27	4.07	6.11

Table 2: FID comparison on FFHQ, LSUN Bedroom and Churches (lower is better).

4.2 Absorbing Diffusion

In this section we analyse the usage of absorbing diffusion for high-resolution image generation, determining how many sampling steps are required to obtain high-quality samples and ablating the components of our approach.

Sampling Speed Our approach applies a diffusion process to a highly compressed image representation, meaning it is already 18× faster to sample from



(a) Non-cherry picked, $\tau = 0.9$, 256×256 LSUN Churches samples.

(b) Non-cherry picked, $\tau = 0.85$, 256×256 FFHQ samples.

(c) Non-cherry picked, $\tau = 0.9$, 256×256 LSUN Bedroom samples.

Fig. 5: Samples from our approach are diverse and high quality.

than DDPM (ours: 3.8s, DDPM: 70s per image on a NVIDIA RTX 2080 Ti). However, since the absorbing diffusion model is trained to approximate $p(\boldsymbol{z}_0|\boldsymbol{z}_t)$ it is possible to speed the sampling process up further by skipping arbitrary numbers of time steps, unmasking multiple latents at once. In Tab. 3 we explore how sample quality is affected using a simple step skipping scheme: evenly skipping a constant number of steps so that the total number of steps meets some fixed computational budget. As expected, FID increases with fewer sampling steps. However, the increase in FID is minor relative to the improvement in sampling speed: our approach achieves similar FID to the equivalent autoregressive model using half the number of steps. With 50 sampling steps, our approach is 88× faster than DDPMs. Using a more sophisticated step selection scheme such as dynamic programming [88], FID could potentially be reduced further.

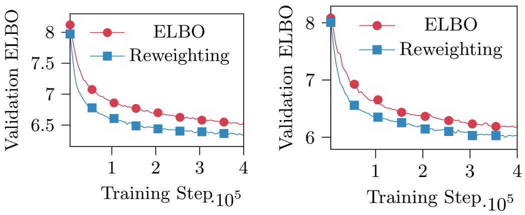
Autoregressive vs Absorbing DDPM Tab. 4 compares the representation ability of our absorbing diffusion model with an autoregressive model, both util-

Steps	50	100	150	200	256
Church	6.86	6.09	5.81	5.68	5.58
Church ($\tau = 0.9$)	4.90	4.40	4.22	4.19	4.07
FFHQ	9.60	7.90	7.53	7.52	7.12
FFHQ ($\tau = 0.9$)	6.87	6.24	6.16	6.14	6.11

Table 3: Our approach allows sampling in much fewer steps with only minor FID increase.

Method	Churches		FFHQ	
	FID ↓	NLL ↓	FID ↓	NLL ↓
*AR	13.23	6.67	9.47	6.65
*Absorbing	11.84	6.41	8.52	6.48
AR	5.93	6.24	8.15	6.18
Absorbing	**5.58**	**6.01**	**7.12**	**5.96**

Table 4: FID and validation latent NLL (in bpd) using the same Transformer. *=Default VQGAN



(a) LSUN Churches

(b) FFHQ

Fig. 6: Models trained with reweighting converge faster than models trained on ELBO.

Modifications	Churches	FFHQ
Default	5.25	3.37
$\lambda_{\max} = 1$	8.67	4.72
DiffAug	5.16	6.57
Both	**2.70**	**3.12**

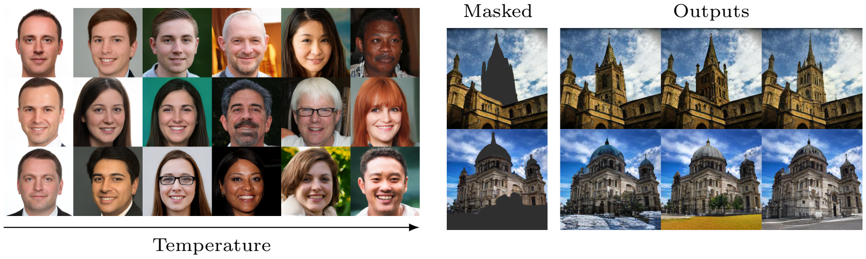
Table 5: Effect of proposed VQGAN changes on FID.

ising exactly the same Transformer architecture, but with the Transformer unconstrained in the diffusion case. On both datasets diffusion achieves lower FID, which is calculated in the image space. Validation NLL is evaluated in latent space (i.e. $-\log p(\boldsymbol{z})$) and again the diffusion model outperforms the autoregressive model despite being trained on a harder task with the same number of parameters, indicating that the diffusion models better approximate the prior distribution. Following previous works, early stopping was used to prevent autoregressive models from overfitting [21,41]; increasing weight decay and dropout in some cases slightly improved validation NLL but caused FID to increase.

Reweighted ELBO In Sec. 3.2 we proposed using a reweighted ELBO when training the diffusion model to reduce gradient variance. We evaluate this in Fig. 6 by comparing validation ELBO (calculated with Eq. 6) during training for models trained directly on ELBO and our reweighting. The models trained on reweighted ELBO converge substantially faster, demonstrating that our reweighting is valid and simplifies optimisation.

4.3 Reconstruction Quality

In Tab. 5 we evaluate the effect of DiffAug [93] and λ limiting on Vector-Quantized image models. While each technique individually can lead to worse FID due to imbalance between the generator and discriminator, we found combining techniques offered the most stability and improved FID across all datasets.



(a) Impact of sampling temperature on diversity. For small temperature changes it is unclear how bias changes.

(b) Our bidirectional approach allows local image editing by targeting regions to be changed (highlighted in grey).

Fig. 7: Evaluation of practical use cases of our proposed generative model.

4.4 Sample Diversity

To improve sample quality, many generative models are sampled using a reduced temperature or by truncating distributions. This is problematic, as these methods amplify any biases in the dataset. We visualise the impact of temperature on sampling from a model trained on FFHQ in Fig. 7a. For very low temperatures the bias the obvious: samples are mostly front-facing white men with brown hair on solid white/black backgrounds. Exactly how the bias changes for more subtle temperature changes is less clear, which is problematic. Practitioners should be aware of this effect and it emphasises the importance of dataset balancing.

4.5 Image Editing

An additional advantage of using a bidirectional diffusion model to model the latent space is that image inpainting is possible. Since autoregressive models are conditioned only on the upper left region of the image, they are unable to edit internal masked image regions in a consistent manner. Diffusion models, on the other hand, allow masked regions to be placed at arbitrary locations. After a region has been highlighted, we mask corresponding latents, identify the starting time step by counting the number of masked latents, then continue the denoising process from that point. Examples of this process can be found in Fig. 7b.

4.6 Limitations

In our experiments we only tested our approach on 256 × 256 datasets; directly scaling to higher resolutions would require more GPU resources. However, future work using more efficient Transformer architectures [39] may alleviate this. Our method outperforms all approaches tested on FID except StyleGAN2 [45]; we find that the primary bottleneck is the Vector-Quantized image model, therefore more research is necessary to improve these discrete representations. Whilst our approach is trained for significantly less time than other approaches such as

StyleGAN2, the stochastic training procedure means that more training steps are required compared to autoregressive approaches. Although when generating extra-large images the large context window made possible by the diffusion model encourages consistency, a reduced temperature is required, reducing diversity.

5 Discussion

While other classes of discrete generative model exist, they are less suitable for Vector-Quantized image modelling than discrete diffusion models: VAEs introduce prior assumptions about the latent space that can be limiting, in particular, continuous spaces may not be appropriate when modelling discrete data [7]; GAN training requires sampling from the generator meaning that gradients must be backpropagated through a discretistion procedure [60]; discrete normalising flows require functions to be invertible, significantly restricting function space [4,37].

Another approach for modelling latent spaces using diffusion models is LSGMs [84], which model continuous latents with SDEs. However, our approach trains more than 15× faster thanks to the efficiency discrete approaches allow. There also exists a variety of different discrete diffusion methods [1,35,79]: ImageBART [21], developed concurrently with this work, models discrete latents using multinomial diffusion with separate autoregressive Transformers per diffusion step leading to slower training, inference, and substantially more parameters than our method. Other concurrent works [10,29,70] which apply diffusion processes to VQGANs are discussed in Appendix D. Also of interest are non-autoregressive discrete methods for translation [24,28,72] and alignment [9,73].

There are a number of avenues that would make for interesting future work based upon the models proposed in this paper: methods that scale diffusion models such as momentum [18], noise schedules [59], cascaded models [34,74] and classifier guidance [16] may yield improved performance. Or, to improve discrete image representations, networks invariant to translation and rotation [43] or other more powerful generative models could be used. Finally, by conditioning on both text and discrete image representations, absorbing diffusion models could allow text-to-image generation and image captioning to be accomplished using a single model with faster run-time than independent approaches [65,67].

6 Conclusion

In this work we proposed a discrete diffusion probabilistic model prior capable of predicting Vector-Quantized image representations in parallel, overcoming the high sampling times, unidirectional nature and overfitting challenges associated with autoregressive priors. Our approach makes no assumptions about the inherent ordering of latents by utilising an unconstrained Transformer architecture. Experimental results demonstrate the ability of our approach to generate diverse, high-quality images, optionally at resolutions exceeding the training samples. Additional work is needed to reduce training times and to efficiently scale our approach to even higher resolutions.

References

1. Austin, J., Johnson, D., Ho, J., Tarlow, D., Berg, R.v.d.: Structured Denoising Diffusion Models in Discrete State-Spaces. arXiv preprint arXiv:2107.03006 (2021) 4, 5, 6, 14, 22
2. Barannikov, S., Trofimov, I., Sotnikov, G., Trimbach, E., Korotin, A., Filippov, A., Burnaev, E.: Manifold Topology Divergence: a Framework for Comparing Data Manifolds. arXiv preprint arXiv:2106.04024 (2021) 9
3. Bengio, Y.: Estimating or propagating gradients through stochastic neurons (2013) 3
4. Berg, R.v.d., Gritsenko, A.A., Dehghani, M., Sønderby, C.K., Salimans, T.: IDF++: Analyzing and Improving Integer Discrete Flows for Lossless Compression. In: International Conference on Learning Representations (2021) 14
5. Bond-Taylor, S., Leach, A., Long, Y., Willcocks, C.G.: Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021). https://doi.org/10.1109/TPAMI.2021.3116668 2, 3
6. Borji, A.: Pros and Cons of GAN Evaluation Measures: New Developments. arXiv preprint arXiv:2103.09396 (2021) 8
7. Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A.M., Jozefowicz, R., Bengio, S.: Generating Sentences from a Continuous Space. arXiv:1511.06349 (2016) 14
8. Brock, A., Donahue, J., Simonyan, K.: Large Scale GAN Training for High Fidelity Natural Image Synthesis. In: International Conference on Learning Representations (2019) 2
9. Chan, W., Saharia, C., Hinton, G., Norouzi, M., Jaitly, N.: Imputer: Sequence Modelling via Imputation and Dynamic Programming. In: International Conference on Machine Learning. pp. 1403–1413. PMLR (2020) 14
10. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: MaskGIT: Masked Generative Image Transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11315–11325 (2022) 14, 23
11. Chen, X., Cohen-Or, D., Chen, B., Mitra, N.J.: Towards a Neural Graphics Pipeline for Controllable Image Generation. In: Computer Graphics Forum. vol. 40, pp. 127–140. Wiley Online Library (2021) 2
12. Child, R.: Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images. In: International Conference on Learning Representations (2021) 2, 9, 10
13. Child, R., Gray, S., Radford, A., Sutskever, I.: Generating Long Sequences with Sparse Transformers. arXiv preprint arXiv:1904.10509 (2019) 2, 5
14. Child, R., Gray, S., Radford, A., Sutskever, I.: Generating Long Sequences with Sparse Transformers. arXiv:1904.10509 (2019) 3
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In: NAACL-HLT (2019) 4, 6
16. Dhariwal, P., Nichol, A.: Diffusion Models Beat GANs on Image Synthesis. Advances in Neural Information Processing Systems **34** (2021) 10, 14
17. Dieleman, S., Oord, A.v.d., Simonyan, K.: The Challenge of Realistic Music Generation: Modelling Raw Audio at Scale. In: Advances in Neural Information Processing Systems. vol. 31 (2018) 7
18. Dockhorn, T., Vahdat, A., Kreis, K.: Score-Based Generative Modeling with Critically-Damped Langevin Diffusion. In: International Conference on Learning Representations (2022) 14

19. Du, Y., Mordatch, I.: Implicit Generation and Generalization in Energy-Based Models. In: Advances in Neural Information Processing Systems. vol. 33 (2019) 4
20. Ericsson, L., Gouk, H., Hospedales, T.M.: How Well Do Self-Supervised Models Transfer? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5414–5423 (2021) 3
21. Esser, P., Rombach, R., Blattmann, A., Ommer, B.: Imagebart: Bidirectional Context with Multinomial Diffusion for Autoregressive Image Synthesis. arXiv preprint arXiv:2108.08827 (2021) 10, 12, 14
22. Esser, P., Rombach, R., Ommer, B.: Taming Transformers for High-Resolution Image Synthesis. arXiv:2012.09841 (2021), `http://arxiv.org/abs/2012.09841` 3, 6, 7, 8, 9, 10, 21
23. Fetty, L., Bylund, M., Kuess, P., Heilemann, G., Nyholm, T., Georg, D., Löfstedt, T.: Latent Space Manipulation for High-Resolution Medical Image Synthesis via the StyleGAN. Zeitschrift für Medizinische Physik **30**(4), 305–314 (2020) 2
24. Ghazvininejad, M., Levy, O., Liu, Y., Zettlemoyer, L.: Mask-Predict: Parallel Decoding of Conditional Masked Language Models. arXiv preprint arXiv:1904.09324 (2019) 14
25. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. In: Advances in Neural Information Processing Systems. vol. 27 (2014) 2
26. Goyal, K., Dyer, C., Berg-Kirkpatrick, T.: Exposing the Implicit Energy Networks behind Masked Language Models via Metropolis–Hastings. arXiv preprint arXiv:2106.02736 (2021) 4
27. Grathwohl, W., Swersky, K., Hashemi, M., Duvenaud, D., Maddison, C.J.: Oops I Took A Gradient: Scalable Sampling for Discrete Distributions. In: International Conference on Machine Learning (2021) 4
28. Gu, J., Wang, C., Zhao, J.: Levenshtein Transformer. Advances in Neural Information Processing Systems **32** (2019) 14
29. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector Quantized Diffusion Model for Text-to-Image Synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10696–10706 (2022) 14, 23
30. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al.: A Survey on Visual Transformer. arXiv preprint arXiv:2012.12556 (2020) 4
31. Hinton, G.E.: Training Products of Experts by Minimizing Contrastive Divergence. Neural Computation **14**(8), 1771–1800 (2002). https://doi.org/10.1162/089976602760128018 4
32. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: Advances in Neural Information Processing Systems. vol. 33 (2020) 4, 6, 10
33. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded Diffusion Models for High Fidelity Image Generation. arXiv preprint arXiv:2106.15282 (2021) 2
34. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded Diffusion Models for High Fidelity Image Generation. Journal of Machine Learning Research **23**(47), 1–33 (2022) 14
35. Hoogeboom, E., Gritsenko, A.A., Bastings, J., Poole, B., van den Berg, R., Salimans, T.: Autoregressive Diffusion Models. In: International Conference on Learning Representations (2022) 14

36. Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., Welling, M.: Argmax Flows and Multinomial Diffusion: Towards Non-Autoregressive Language Models. arXiv preprint arXiv:2102.05379 (2021) 4, 6
37. Hoogeboom, E., Peters, J., van den Berg, R., Welling, M.: Integer Discrete Flows and Lossless Compression. In: Advances in Neural Information Processing Systems. vol. 32 (2019) 14
38. Hudson, D.A., Zitnick, C.L.: Generative Adversarial Transformers. Proceedings of the 38th International Conference on Machine Learning, ICML (2021) 9
39. Jaegle, A., Borgeaud, S., Alayrac, J.B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., et al.: Perceiver IO: A General Architecture for Structured Inputs & Outputs. arXiv preprint arXiv:2107.14795 (2021) 13
40. Jang, E., Gu, S., Poole, B.: Categorical Reparameterization with Gumbel-Softmax. In: International Conference on Learning Representations (2017) 7
41. Jun, H., Child, R., Chen, M., Schulman, J., Ramesh, A., Radford, A., Sutskever, I.: Distribution Augmentation for Generative Modeling. In: ICML (2020) 3, 7, 12
42. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive Growing of GANs for Improved Quality, Stability, and Variation. In: International Conference on Learning Representations (2018) 2, 9, 10
43. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-Free Generative Adversarial Networks. arXiv preprint arXiv:2106.12423 (2021) 14
44. Karras, T., Laine, S., Aila, T.: A Style-Based Generator Architecture for Generative Adversarial Networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019) 8, 9
45. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and Improving the Image Quality of StyleGAN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020) 9, 10, 13
46. Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. In: International Conference on Machine Learning (2020) 5
47. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. arXiv preprint arXiv:1412.6980 (2014) 8, 21
48. Kingma, D.P., Salimans, T., Poole, B., Ho, J.: Variational Diffusion Models. arXiv preprint arXiv:2107.00630 (2021) 3
49. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. In: International Conference on Learning Representations (2014) 2
50. Kingma, D.P., Dhariwal, P.: Glow: Generative Flow with Invertible 1x1 Convolutions. In: Advances in Neural Information Processing Systems. vol. 31 (2018) 2
51. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved Precision and Recall Metric for Assessing Generative Models. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32 (2019) 9
52. Lin, J., Zhang, R., Ganz, F., Han, S., Zhu, J.Y.: Anycost GANs for Interactive Image Synthesis and Editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14986–14996 (2021) 2
53. Liu, B., Zhu, Y., Song, K., Elgammal, A.: Towards Faster and Stabilized GAN Training for High-fidelity Few-shot Image Synthesis. In: International Conference on Learning Representations (2021) 2

54. Maddison, C.J., Mnih, A., Teh, Y.W.: The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In: International Conference on Learning Representations (2017) 7
55. Menick, J., Kalchbrenner, N.: Generating High Fidelity Images with Subscale Pixel Networks and Multidimensional Upscaling. In: International Conference on Learning Representations (2019) 2
56. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral Normalization for Generative Adversarial Networks. In: International Conference on Learning Representations (2018) 2
57. Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J.: Reliable Fidelity and Diversity Metrics for Generative Models. In: International Conference on Machine Learning. pp. 7176–7185 (2020) 9, 22
58. Nash, C., Menick, J., Dieleman, S., Battaglia, P.W.: Generating Images with Sparse Representations. arXiv preprint arXiv:2103.03841 (2021) 9, 10, 22
59. Nichol, A.Q., Dhariwal, P.: Improved Denoising Diffusion Probabilistic Models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021) 14, 23
60. Nie, W., Narodytska, N., Patel, A.: RelGAN: Relational Generative Adversarial Networks for Text Generation. In: International Conference on Learning Representations (2019) 14
61. Obukhov, A., Seitzer, M., Wu, P.W., Zhydenko, S., Kyl, J., Lin, E.Y.J.: High-fidelity performance metrics for generative models in pytorch (2020). https://doi.org/10.5281/zenodo.4957738, `https://github.com/toshas/torch-fidelity`, version: 0.3.0, DOI: 10.5281/zenodo.4957738 10
62. van den Oord, A., Kalchbrenner, N., Espeholt, L., kavukcuoglu, k., Vinyals, O., Graves, A.: Conditional Image Generation with PixelCNN Decoders. NeurIPS 29 (2016) 3
63. van den Oord, A., Vinyals, O., kavukcuoglu, k.: Neural Discrete Representation Learning. NeurIPS 30 (2017) 2, 3, 8
64. Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., Tran, D.: Image Transformer. In: ICML (2018) 2
65. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning Transferable Visual Models From Natural Language Supervision. arXiv preprint arXiv:2103.00020 (2021) 14
66. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language Models are Unsupervised Multitask Learners (2019) 8, 21
67. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-Shot Text-to-Image Generation. arXiv preprint arXiv:2102.12092 (2021) 7, 14
68. Razavi, A., van den Oord, A., Vinyals, O.: Generating Diverse High-Fidelity Images with VQ-VAE-2. NeurIPS 32 (2019) 7, 9, 10
69. Reed, S., Oord, A., Kalchbrenner, N., Colmenarejo, S.G., Wang, Z., Chen, Y., Belov, D., Freitas, N.: Parallel Multiscale Autoregressive Density Estimation. In: International Conference on Machine Learning (2017) 2, 5
70. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., 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) 14, 23
71. Roy, A., Saffar, M., Vaswani, A., Grangier, D.: Efficient Content-Based Sparse Attention with Routing Transformers. Transactions of the Association for Computational Linguistics **9**, 53–68 (2021) 3

72. Ruis, L., Stern, M., Proskurnia, J., Chan, W.: Insertion-Deletion Transformer. arXiv preprint arXiv:2001.05540 (2020) 14
73. Saharia, C., Chan, W., Saxena, S., Norouzi, M.: Non-Autoregressive Machine Translation with Latent Alignments. arXiv preprint arXiv:2004.07437 (2020) 14
74. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image Super-Resolution via Iterative Refinement. arXiv preprint arXiv:2104.07636 (2021) 14
75. Sajjadi, M.S., Bachem, O., Lucic, M., Bousquet, O., Gelly, S.: Assessing Generative Models via Precision and Recall. In: Advances in Neural Information Processing Systems. vol. 31 (2018) 9
76. Salimans, T., Karpathy, A., Chen, X., Kingma, D.P.: PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications. ICLR (2017) 3
77. Sasaki, H., Willcocks, C.G., Breckon, T.P.: UNIT-DDPM: UNpaired Image Translation with Denoising Diffusion Probabilistic Models. arXiv preprint arXiv:2104.05358 (2021) 2
78. Sauer, A., Chitta, K., Müller, J., Geiger, A.: Projected GANs Converge Faster. Advances in Neural Information Processing Systems **34**, 17480–17492 (2021) 9, 10
79. Savinov, N., Chung, J., Binkowski, M., Elsen, E., van den Oord, A.: Step-unrolled Denoising Autoencoders for Text Generation. In: International Conference on Learning Representations (2022) 14
80. Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In: International Conference on Machine Learning (2015) 4
81. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-Based Generative Modeling through Stochastic Differential Equations. In: International Conference on Learning Representations (2021) 2
82. Theis, L., Oord, A.v.d., Bethge, M.: A note on the evaluation of generative models. arXiv:1511.01844 (2016) 8
83. Tsitsulin, A., Munkhoeva, M., Mottin, D., Karras, P., Bronstein, A., Oseledets, I., Müller, E.: The Shape of Data: Intrinsic Distance for Data Distributions. In: International Conference on Learning Representations (2020) 8, 9
84. Vahdat, A., Kreis, K., Kautz, J.: Score-Based Generative Modeling in Latent Space. Advances in Neural Information Processing Systems **34** (2021) 2, 14
85. Van Den Oord, A., Kalchbrenner, N., Kavukcuoglu, K.: Pixel Recurrent Neural Networks. In: ICML (2016) 3
86. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, ., Polosukhin, I.: Attention is All you Need. In: Advances in Neural Information Processing Systems. vol. 30 (2017) 2, 4, 6, 8
87. Wang, A., Cho, K.: BERT has a Mouth, and it Must Speak: BERT as a Markov Random Field Language Model. In: NeuralGen (2019) 4
88. Watson, D., Ho, J., Norouzi, M., Chan, W.: Learning to Efficiently Sample from Diffusion Probabilistic Models. arXiv preprint arXiv:2106.03802 (2021) 11
89. Xiao, Z., Kreis, K., Kautz, J., Vahdat, A.: VAEBM: A Symbiosis between Variational Autoencoders and Energy-based Models. In: International Conference on Learning Representations (2021) 2
90. Yu, F., Zhang, Y., Song, S., Seff, A., Xiao, J.: Lsun: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. arXiv preprint arXiv:1506.03365 (2015) 8
91. Yu, N., Barnes, C., Shechtman, E., Amirghodsi, S., Lukac, M.: Texture Mixer: A Network for Controllable Synthesis and Interpolation of Texture. In: Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12164–12173 (2019) 2

92. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 3, 24
93. Zhao, S., Liu, Z., Lin, J., Zhu, J.Y., Han, S.: Differentiable Augmentation for Data-Efficient GAN Training. In: Advances in Neural Information Processing Systems. vol. 33 (2020) 7, 12

Supplementary Material

The supplementary material for this work is divided into the following sections: Appendix A describes the architectures and hyperparameters for the experiments presented in the main paper; Appendix B illustrates the connection between our proposed ELBO reweighting and the true ELBO; Appendix C contains extra FID comparisons; Appendix D compares our approach with concurrent works; Appendix E gives nearest neighbour examples to demonstrate generalisation; and finally, Appendix F contains additional samples at resolutions higher than the training data.

A Implementation Details

We perform all experiments on a single NVIDIA RTX 2080 Ti with 11GB of VRAM using automatic mixed precision when possible. As mentioned in the main paper, we use the same VQGAN architecture as used by Esser et al. [22] which for 256×256 images downsamples to features of size $16 \times 16 \times 256$, and quantizes using a codebook with 1024 entries. Attention layers are applied within both the encoder and decoder on the lowest resolutions to aggregate context across the entire image. Models are optimised using the Adam optimiser [47] using a batch size of 4 and learning rate of 1.8×10^{-5}. For the differentiable augmentations we randomly change the brightness, saturation, and contrast, as well as randomly translate images. The datasets we use are both publically accessible, with FFHQ availble under the Creative Commons BY 4.0 licence. LSUN models are trained for 2.2M steps and the FFHQ model for 1.4M steps.

For the absorbing diffusion model we use a scaled down 80M parameter version of GPT-2 [66] consisting of 24 layers, where each attention layer has 8 heads, each 64D. The same architecture is used for experiments with the autoregressive model. Autoregressive models' training are stopped based on the best validation loss. We also stop training the absorbing diffusion models based on validation ELBO, however, on the LSUN datasets we found that it always improved or remained consistent throughout training so each model was trained for 2M steps.

Codebook Collapse One issue with vector quantized methods is codebook collapse, where some codes fall out of use which limits the potential expressivity of the model. We found this to occur across all datasets with often a fraction of the codes in use. We experimented with different quantization schemes such as gumbel softmax, different initalisation schemes such as k-means, and 'code recycling', where codes out of use are reset to an in use code. In all of these cases, we found the reconstruction quality to be comparable or worse so stuck with the argmax quantisation scheme used by Esser et al. [22].

Precision, Recall, Density, and Coverage To compute these measures we use the official code releases and pretrained weights in all cases except Taming Transformers on the LSUN datasets where weights were not available; in this case we reproduced results as close as possible with the hardware available, training the VQGANs and autoregressive models with the same hyperparameters used for the rest of our experiments. Following Nash et al. [58] we use the standard 2048D InceptionV3 features, which are also used to compute FID, $k = 3$ nearest neighbours, and 50k samples, and use the code provided by Naeem et al. [57].

B Reweighted ELBO

In Sec. 3.2 we propose re-weighting the ELBO of the absorbing diffusion model so that the individual loss at each time step is multiplied by $\frac{T-t+1}{T}$ rather than $1/t$. In this section we justify the correctness of this re-weighting by showing it is equivalent to minimising the difference to a forward process that does not have access to $\boldsymbol{x}_t$. As such, the loss takes into account the difficulty of denoising steps and re-weights them down accordingly. This derivation is based on the true ELBO derived by [1]. The loss at time step t can be written as

$$\begin{aligned}\mathcal{L}_t &= D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0) \parallel p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) \\ &= \sum_i \sum_j q([\boldsymbol{x}_{t-1}]_{i,j}|\boldsymbol{x}_0) \log \frac{q([\boldsymbol{x}_{t-1}]_{i,j}|\boldsymbol{x}_0)}{p([\boldsymbol{x}_{t-1}]_{i,j}|\boldsymbol{x}_t)},\end{aligned} \tag{11}$$

where the first summation sums over latent coordinates i, and the second summation sums over the probabilities of each code j. For the absorbing diffusion case where tokens in $\boldsymbol{x}_t$ are masked independently and uniformly with probability $\frac{t}{T}$, this posterior is defined as

$$q([\boldsymbol{x}_{t-1}]_i = a|\boldsymbol{x}_0) = \begin{cases} 1 - \frac{t-1}{T}, & \text{if } a = [\boldsymbol{x}_0]_i \text{ and } [\boldsymbol{x}_t]_i = m. \\ \frac{t-1}{T}, & \text{if } a = m \text{ and } [\boldsymbol{x}_t]_i = m. \\ 1, & \text{if } a = [\boldsymbol{x}_0]_i \text{ and } [\boldsymbol{x}_t]_i = [\boldsymbol{x}_0]_i. \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

The reverse process remains defined in the same way as the standard reverse process:

$$p([\boldsymbol{x}_{t-1}]_i = a|\boldsymbol{x}_t) = \begin{cases} \frac{1}{t} p_\theta([\boldsymbol{x}_0]_i|\boldsymbol{x}_t), & \text{if } a = [\boldsymbol{x}_0]_i \text{ and } [\boldsymbol{x}_t]_i = m. \\ 1 - \frac{1}{t}, & \text{if } a = m \text{ and } [\boldsymbol{x}_t]_i = m. \\ 1, & \text{if } a = [\boldsymbol{x}_0]_i \text{ and } [\boldsymbol{x}_t]_i = [\boldsymbol{x}_0]_i. \end{cases} \tag{13}$$

Substituting these definitions into Eq. (11), the loss can be simplified to Eq. (14); by extracting the constants into a single term out of the sum, C, the loss can be further simplified to obtain Eq. (15), which is equivalent to our proposed reweighted ELBO Eq. (11),

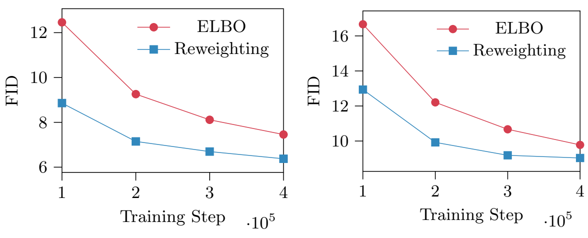
$$\mathcal{L}_t = \sum_i \left[1 \log \frac{1}{1} + \frac{t-1}{T} \log \frac{\frac{t-1}{T}}{1 - \frac{1}{t}} + \left(1 - \frac{t-1}{T} \log \frac{1 - \frac{t-1}{T}}{\frac{1}{t} p_\theta([\boldsymbol{x}_0]_i|\boldsymbol{x}_t)} \right) \right], \tag{14}$$

$$= C - \sum_i \left[\frac{T-t+1}{T} \log p_\theta([\boldsymbol{x}_0]_i|\boldsymbol{x}_t) \right]. \tag{15}$$

C Additional Comparisons

In Fig. 6 we demonstrated that models trained using our proposed ELBO reweighting converge faster in terms of validation ELBO. To further substantiate this and show that improvements extend to sample quality we compare models trained directly on ELBO and our reweighting in terms of FID in Fig. 8. The same trend is observed, with the models trained on the reweighting converging faster.

Since a key property of DDPMs is that sampling times can be reduced by skipping time steps, in Fig. 9 we compare FID scores for various numbers of sampling steps with a continuous DDPM applied in pixel space [59]. We find that our approach using a discrete DDPM and Vector-Quantized image model degrades in performance at a slower rate than the continuous DDPM likely due to the reduced dimensionality, allowing sampling with fewer steps while maintaining quality. In both cases, the performance for very low numbers of sampling steps could potentially be improved with more sophisticated step selection schemes.



(a) LSUN Churches

(b) FFHQ

Fig. 8: Models trained with our reweighted ELBO converge faster than models trained directly on ELBO.

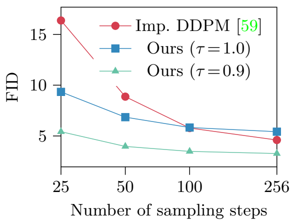


Fig. 9: FID vs number of sampling steps on LSUN Bedroom.

D Concurrent Works

Concurrent with our work, a number of similar approaches independently proposed using diffusion-like models to model VQGAN latents, these approaches are complementary to ours and distinct in a number of ways. VQ-Diffusion [29] use a combination of multinomial and absorbing diffusion to encourage the model to focus less on mask tokens. This, however, requires the use of an additional auxilliary objective function to improve stability, and in practice our approach achieves lower FID on the only shared dataset, FFHQ. MaskGIT [10] models discrete latents by learning to unmask tokens using a similar training scheme to ours; during sampling, tokens are unmasked based on the model's confidence. This approach allows sampling in very few steps, but the lack of theoretical justification makes it unclear how representative samples are. Latent Diffusion [70] relaxes the discrete assumption, using continuous diffusion parameterised by a convolutional U-Net to model latents of greater spatial size, but with lower dimensional codes. Both compressing spatially/depth-wise and discrete/continuous diffusion come with different trade-offs such as sampling time.

E Nearest Neighbours

When training generative models, being able to detect overfitting is key to ensure the data distribution is well modelled. Overfitting is not detected by popular metrics such as FID, making overfitting difficult to identify in approaches such as GANs. With our approach we are able to approximate the ELBO on a validation set making it simple to prevent overfitting. In this section we demonstrate that our approach is not overfit by providing nearest neighbour images from the training dataset to samples from our model, measured using LPIPS [92].

F Additional Samples

Fig. 13 contains unconditional samples with resolutions larger than observed in the training data from a model trained on LSUN Bedroom.

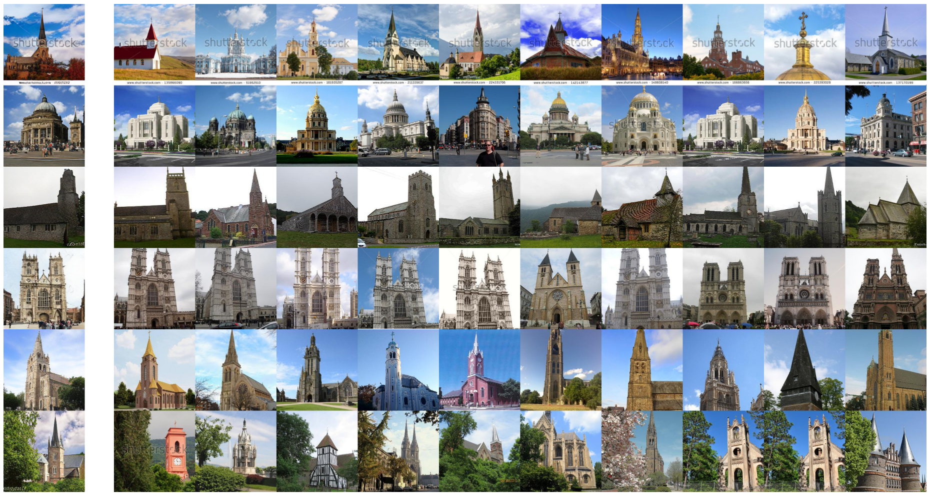


Fig. 10: Nearest neighbours for a model trained on LSUN Churches based on LPIPS distance. The left column contains samples from our model and the right column contains the nearest neighbours in the training set (increasing in distance from left to right).

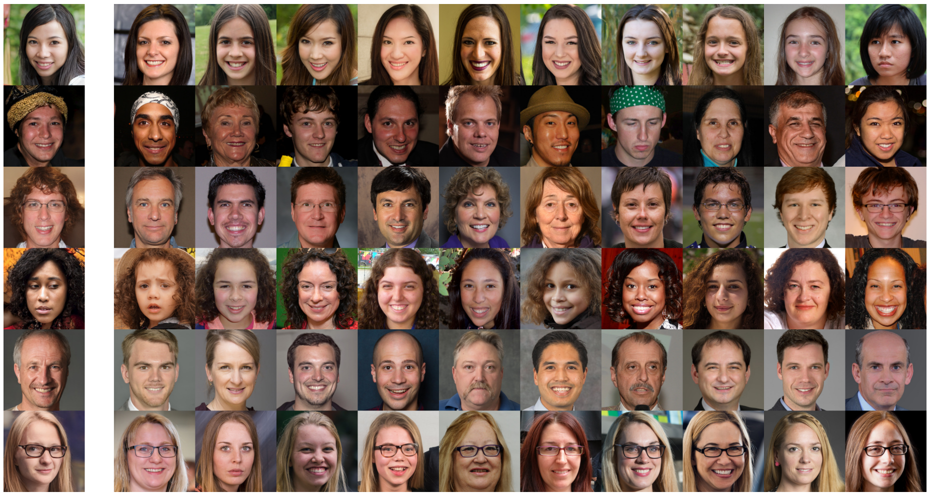


Fig. 11: Nearest neighbours for a model trained on FFHQ based on LPIPS distance. The left column contains samples from our model and the right column contains the nearest neighbours in the training set (increasing in distance from left to right).

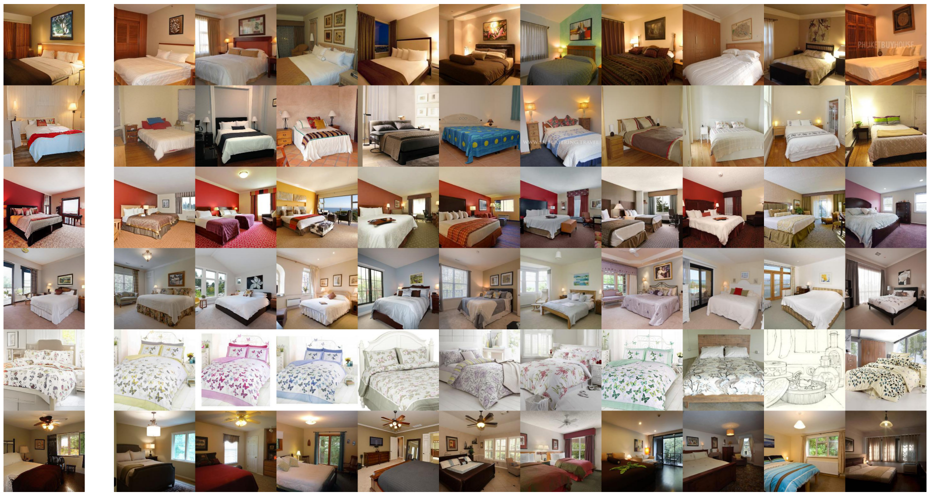


Fig. 12: Nearest neighbours for a model trained on LSUN Bedroom based on LPIPS distance. The left column contains samples from our model and the right column contains the nearest neighbours in the training set (increasing in distance from left to right).

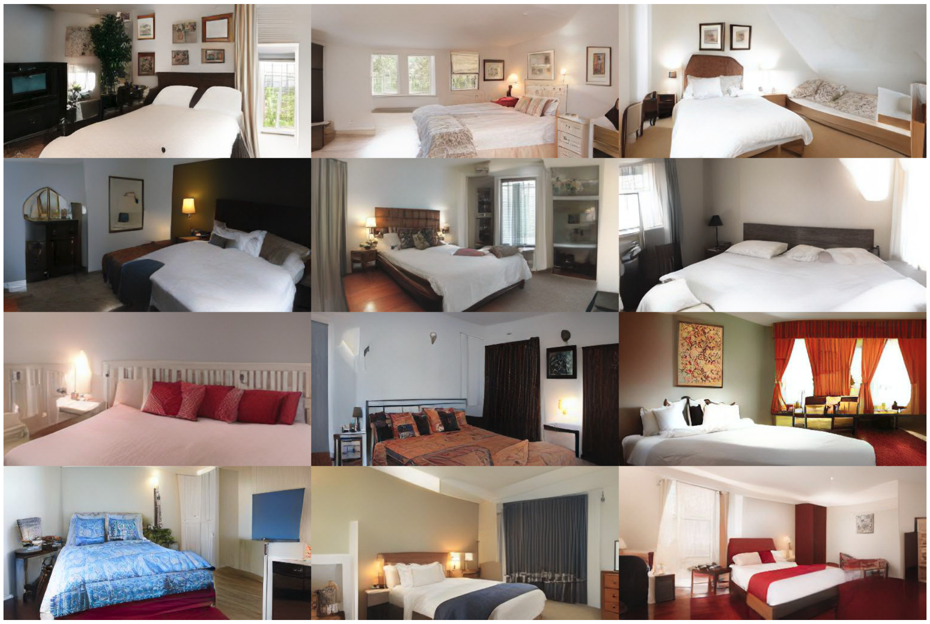


Fig. 13: Unconditional samples from a model trained on LSUN Bedroom larger than images in the training dataset.