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

Sam Bond-Taylor*, Peter Hessey*, Hiroshi Sasaki, Toby P. Breckon, and Chris G. Willcocks

Durham University

Project Page

Introduction

A parallel prediction approach that allows **fast generation** of high-res Vector-Quantized images:

- SOTA quantitative sample quality scores.
- Sample resolutions higher than training data.
- Global context allows samples to be edited.

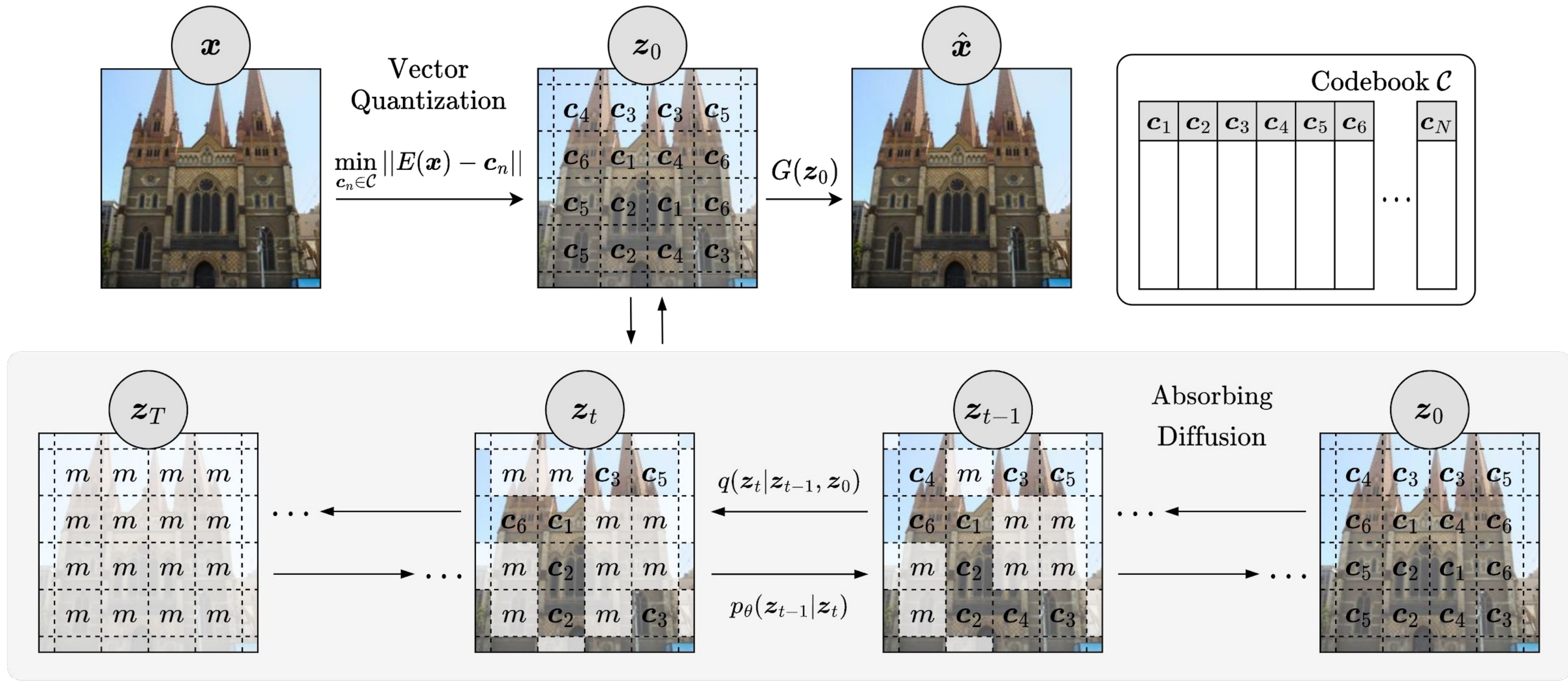
Method

Discrete latents **z** obtained using a VQ-VAE are modelled with discrete **absorbing diffusion**:

$$p_\theta(\boldsymbol{z}_{0:T}) = p_\theta(\boldsymbol{z}_T)\prod_{t=1}^{T} p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t)$$

Elements of **z** are randomly masked and an **unconstrained Transformer** learns to denoise the data by optimising a carefully reweighted ELBO designed to improve convergence rates:

$$\sum_{t=1}^{T} \frac{T-t+1}{T}\mathbb{E}_{q(\boldsymbol{z}_t|\boldsymbol{z}_0)}\Big[\sum_{[\boldsymbol{z}_t]_i=m} \log p_\theta([\boldsymbol{z}_0]_i|\boldsymbol{z}_t)\Big]$$



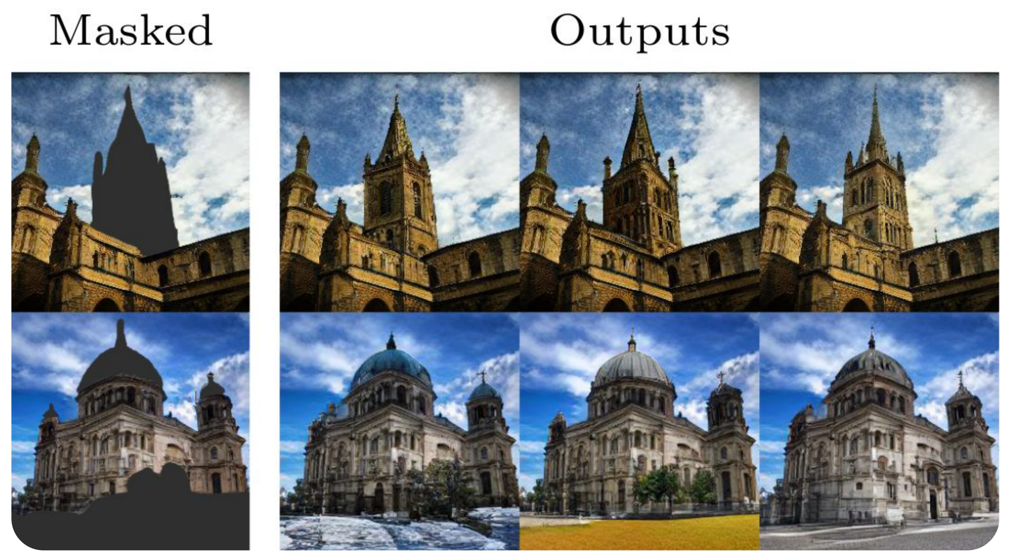
A discrete absorbing diffusion process destroys latents by slowly masking out tokens over many steps, similar to masked language models like BERT.

Sampling at Higher Resolutions

Globally consistent images at **higher resolutions** than the training data can be generated by aggregating multiple context windows:

Image Editing

Our bidirectional approach with **global context** allows internal image regions to be edited by masking those areas (highlighted in grey).



Quantitative Results

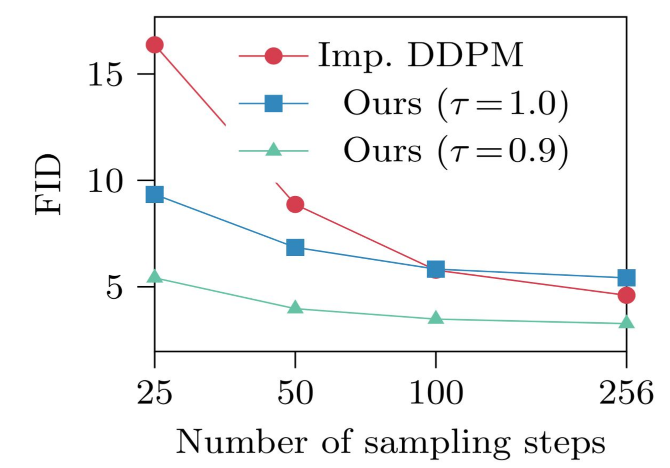
Our models achieve competitive FID scores on LSUN and FFHQ. Evaluating our models using Precision, Recall, Density and Coverage (PRDC) metrics further demonstrates **SOTA** results:

	Churches			Bedroom			FFHQ		
Model	FID↓	D↑	C↑	FID↓	D↑	C↑	FID↓	D↑	C↑
TT	7.81	**1.08**	0.60	6.35	1.15	0.75	9.6	0.89	0.50
ImageBART	7.32	-	-	5.51	-	-	9.57	-	-
StyleGAN2	3.85	0.83	0.68	2.35	-	-	3.80	1.12	0.80
ProjGAN	**1.59**	0.65	0.64	**1.52**	0.90	0.79	**3.39**	0.98	0.77
Ours	4.07	1.07	**0.74**	3.27	**1.51**	**0.83**	6.11	**1.20**	**0.80**

Sampling Speed

Faster sampling can be achieved by predicting tokens in parallel with only small FID change:

Steps	Church	FFHQ
50	4.90	6.87
100	4.40	6.24
150	4.22	6.16
200	4.19	6.14
256	4.07	6.11



Summary

Using a discrete absorbing diffusion model parameterised by an unconstrained Transformer to model VQ-VAE representations we achieve faster sampling with higher visual quality.

Github repository with trained models at *https://samb-t.github.io/unleashing-transformers*

*Authors contributed equally

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

Sam Bond-Taylor*, Peter Hessey*, Hiroshi Sasaki, Toby P. Breckon, and Chris G. Willcocks

Durham University

Project Page

Introduction

We propose a parallel prediction approach for training and sampling generative models capable of generating high-resolution images:

- **Faster sampling** than autoregressive priors
- Inpainting permitted by the bidirectionality
- Sample resolutions higher than training data

Method

Discrete latents $\mathbf{z}$ obtained using a VQVAE are modelled with discrete absorbing diffusion

$$p_\theta(\boldsymbol{x}_{0:T}) = p_\theta(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$$

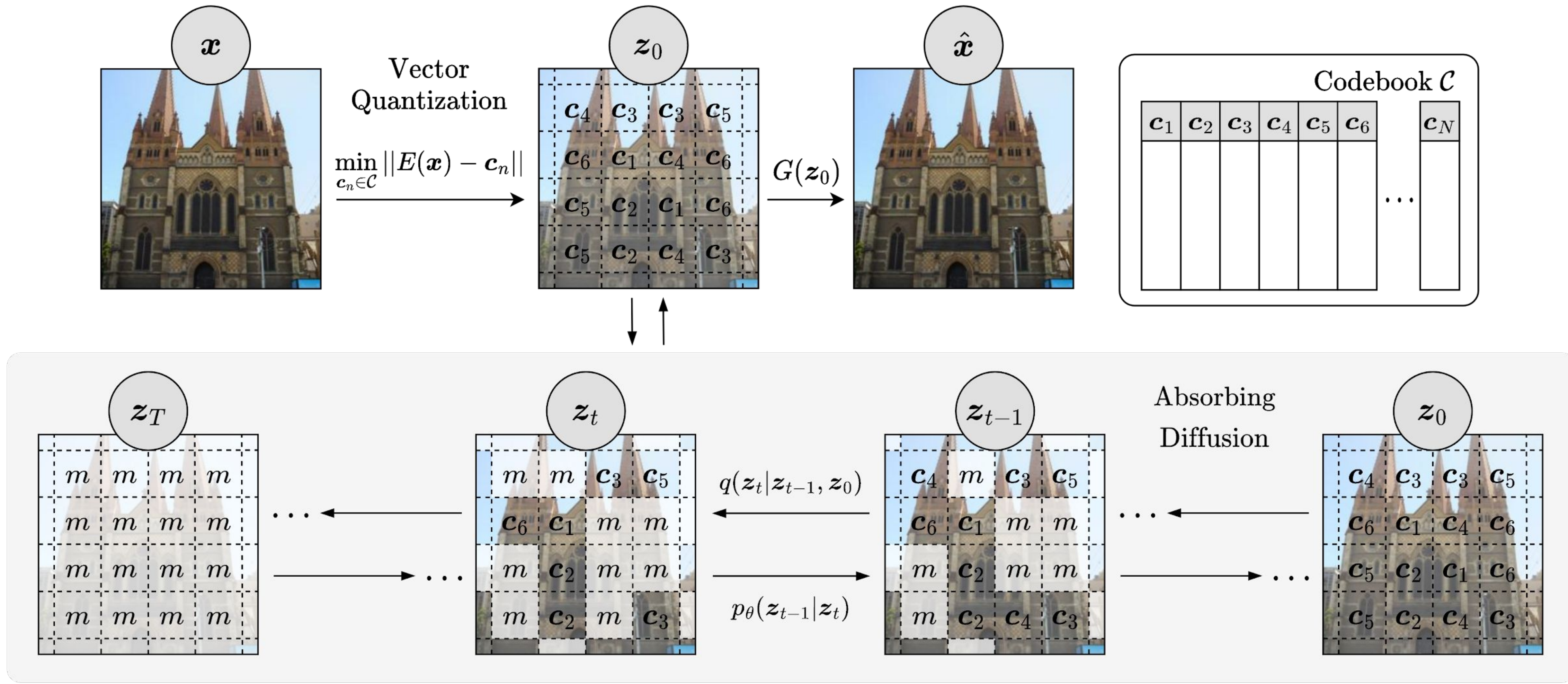
Elements of $\mathbf{z}$ are randomly masked and an **unconstrained Transformer** learns to denoise the data by optimising a carefully reweighted ELBO designed to improve convergence rates

$$\sum_{t=1}^{T} \frac{T-t+1}{T} \mathbb{E}_{q(\boldsymbol{z}_t|\boldsymbol{z}_0)} \Big[\sum_{[\boldsymbol{z}_t]_i = m} \log p_\theta([\boldsymbol{z}_0]_i|\boldsymbol{z}_t) \Big]$$

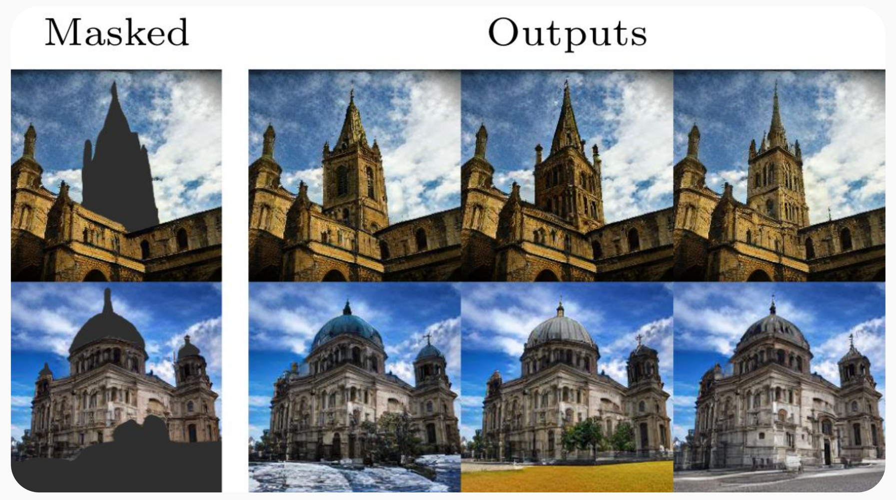
Sampling Speed

Sampling times can be significantly reduced by skipping steps with only small FID change

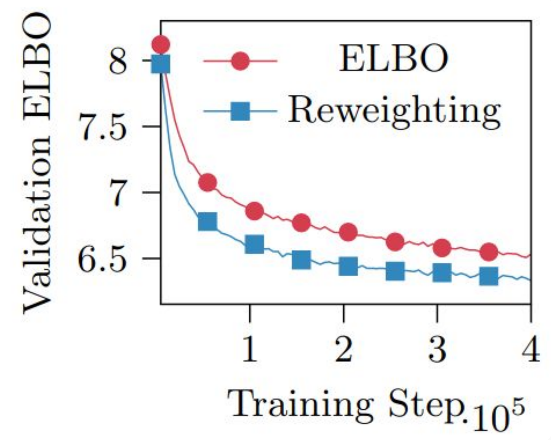
Steps	50	100	150	200	256
Church	4.90	4.40	4.22	4.19	4.07
FFHQ	6.87	6.24	6.16	6.14	6.11



Inpainting



Reweighted ELBO



Improved VQGAN

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**

Samples

Quantitative Results

We achieve highly-competitive FID scores on LSUN and FFHQ. Evaluating our models using Precision, Recall, Density and Coverage (PRDC) metrics further demonstrates **SOTA** results:

	Churches			Bedroom			FFHQ		
Model	FID ↓	D ↑	C ↑	FID ↓	D ↑	C ↑	FID ↓	D ↑	C ↑
TT	7.81	**1.08**	0.60	6.35	1.15	0.75	9.6	0.89	0.50
ImageBART	7.32	-	-	5.51	-	-	9.57	-	-
StyleGAN2	3.85	0.83	0.68	2.35	-	-	3.80	1.12	0.80
ProjGAN	**1.59**	0.65	0.64	**1.52**	0.90	0.79	**3.39**	0.98	0.77
Ours	4.07	1.07	**0.74**	3.27	**1.51**	**0.83**	6.11	**1.20**	**0.80**

Higher Resolutions

Globally consistent images at resolutions greater than the training data can be generated by aggregating multiple context windows

Summary

We improve the sampling speed and quality of VQVAE based generative models by using a discrete diffusion process powered by an unconstrained Transformer.

Github repository with trained models at ***https://samb-t.github.io/unleashing-transformers***

*Authors contributed equally

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

Durham University

Sam Bond-Taylor*, Peter Hessey*, Hiroshi Sasaki, Toby P. Breckon, and Chris G. Willcocks

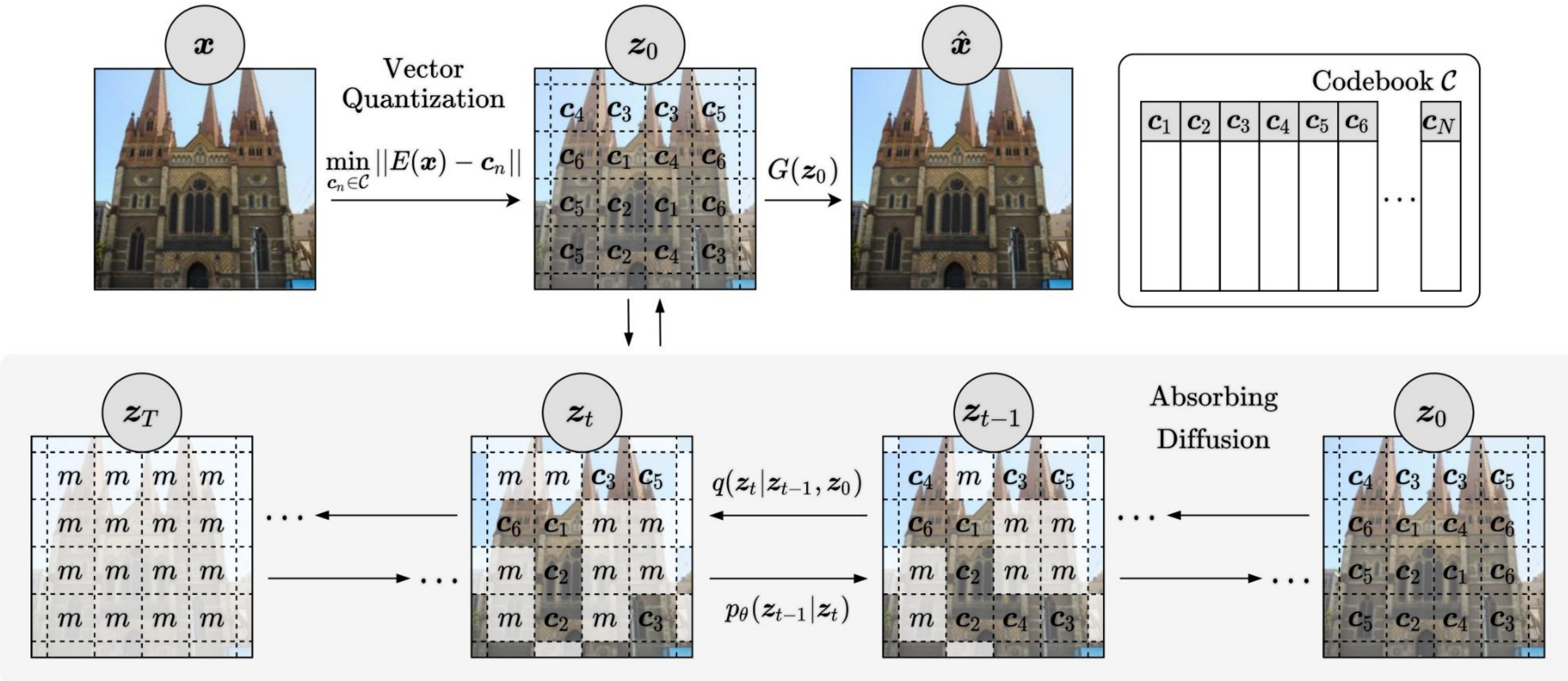
Approach

We propose a parallel prediction approach for the training and sampling of generative models capable of generating high-resolution images. A two-stage training process is utilised: firstly we model a discrete and highly-compressed latent space, before then modelling a prior on this discrete latent space through a powerful discrete diffusion technique inspired by Masked Language Models.

Reweighted ELBO

To increase the speed of convergence when training the discrete diffusion prior, we introduce the following novel ELBO reweighting based upon the unique properties of the techniques applied:

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

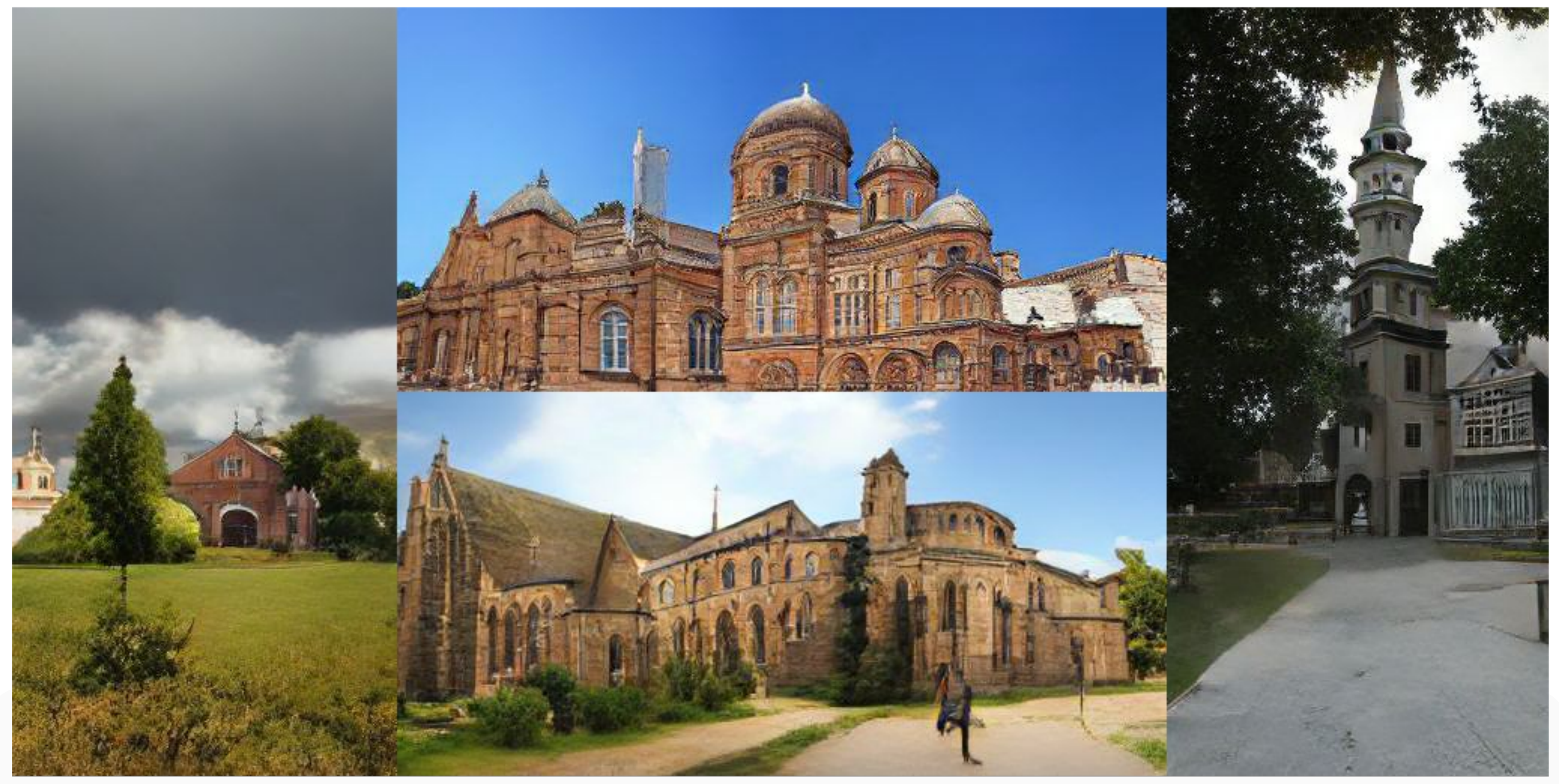


Quantitative Results

We achieve highly-competitive FID scores on LSUN and FFHQ. Evaluating our models using Precision, Recall, Density and Coverage (PRDC) metrics further demonstrates SOTA results:

	LSUN Churches				FFHQ			
Model	P ↑	R ↑	D ↑	C ↑	P ↑	R ↑	D ↑	C ↑
TT	0.67	0.29	1.08	0.60	0.64	0.29	0.89	0.5
ProjGAN	0.56	**0.53**	0.65	0.64	0.66	0.46	0.98	0.77
Ours ($\tau = 1.0$)	0.70	0.42	**1.12**	0.73	0.69	0.48	1.06	0.77
Ours ($\tau = 0.9$)	**0.71**	0.45	1.07	**0.74**	**0.73**	**0.48**	**1.20**	**0.80**

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



Summary

....

Github repository with trained models at ***https://samb-t@github.io/unleashing-transformers***

*Authors contributed equally

Project Page