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**:

$$p_\theta(z_{0:T}) = p_\theta(z_T) \prod_{t=1}^{T} p_\theta(z_{t-1}|z_t)$$

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:

$$\sum_{t=1}^{T} \frac{T-t+1}{T} \mathbb{E}_{q(z_t|z_0)} \Big[\sum_{[z_t]_i=m} \log p_\theta([z_0]_i|z_t) \Big]$$
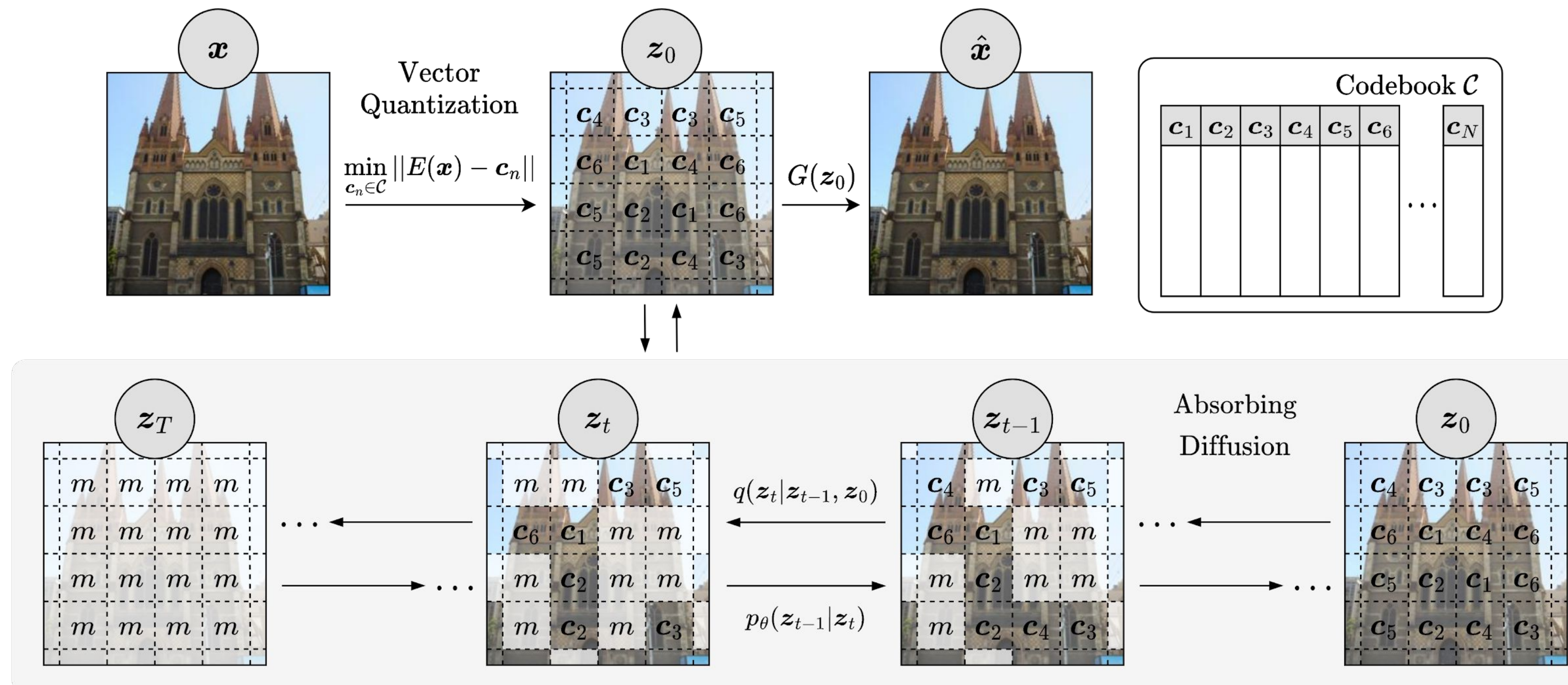
A discrete absorbing diffusion process destroys latents by slowly masking out tokens over many steps, similar to masked language models like BERT.
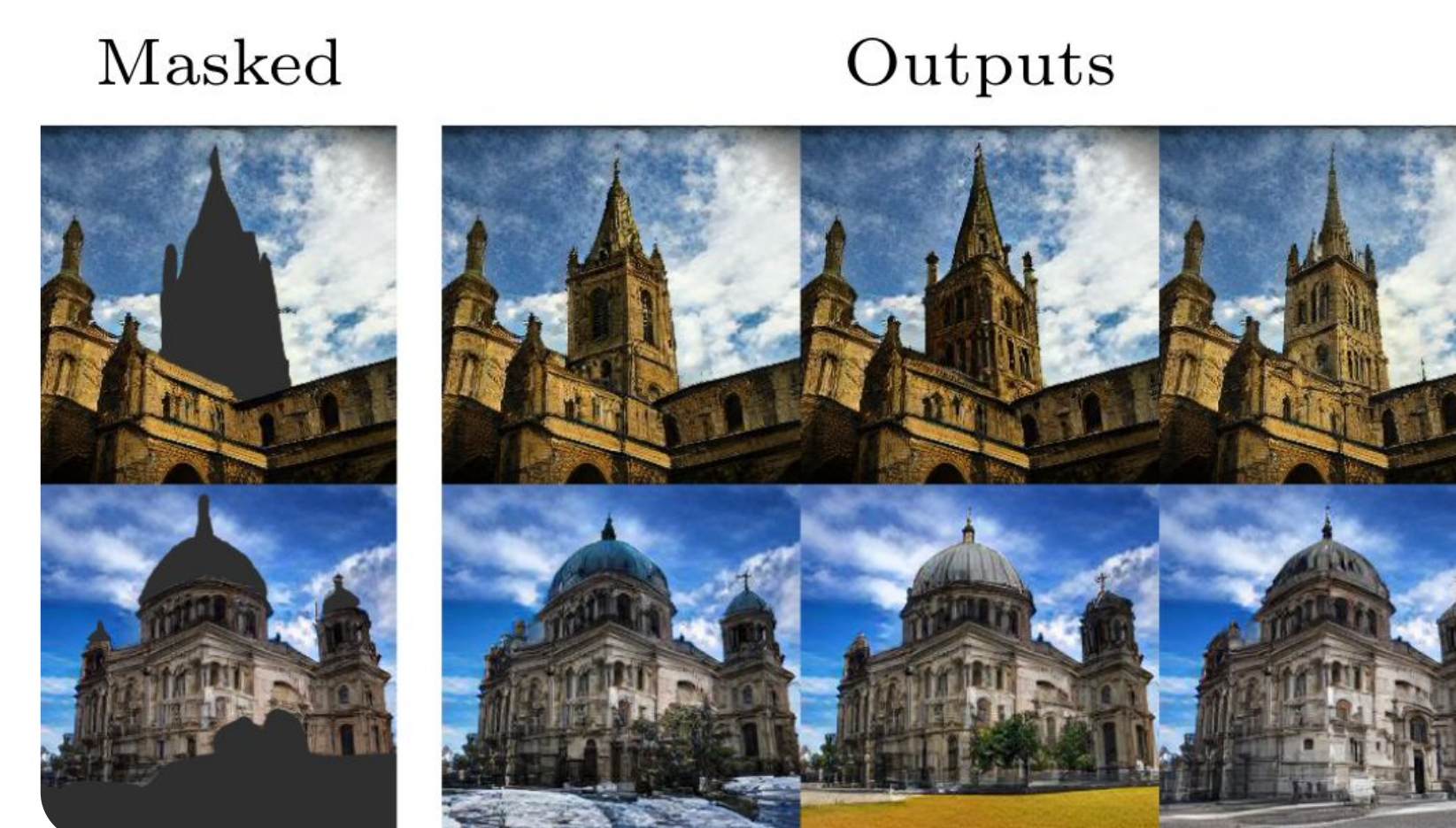
Sampling at Higher Resolutions

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).


Masked · Outputs

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:

Model	Churches FID↓	D↑	C↑	Bedroom FID↓	D↑	C↑	FFHQ 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
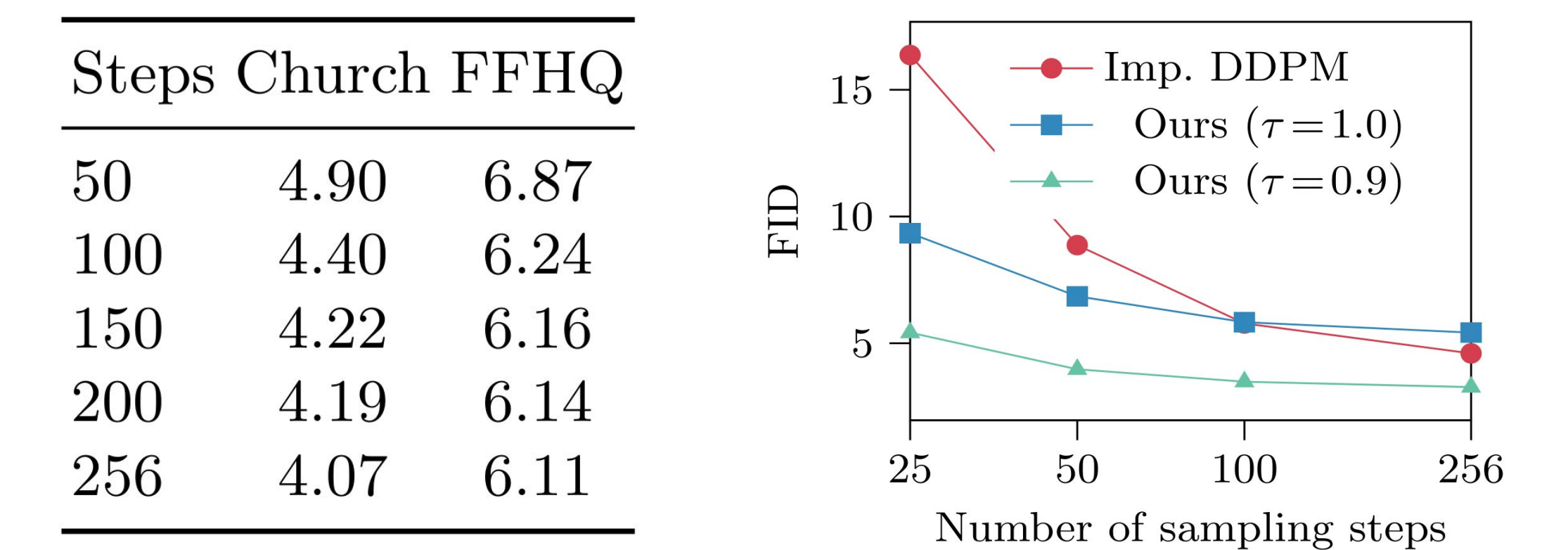
Summary

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