A Appendix

A.1 Limitations and Broader Impact

Limitations While our ConPreDiff boosts performance of both discrete and continuous diffusion models without introducing additional parameters in model inference, our models still have more trainable parameters than other types of generative models, e.g GANs. Furthermore, we note the long sampling times of both and compared to single step generative approaches like GANs or VAEs. However, this drawback is inherited from the underlying model class and is not a property of our context prediction approach. Neighborhood context decoding is fast and incurs negligible computational overhead in training stage. For future work, we will try to find more intrinsic information to preserve for improving existing point-wise denoising diffusion models, and extend to more challenging tasks like text-to-3D and text-to-video generation.

Broader Impact Recent generative image models enable creative applications and autonomous media creation, but can also be viewed as a dual-use technology with negative implications. In this paper, we use human face datasets only for evaluating the image inpainting performance of our method, and our method is not intended to create content that is used to mislead or deceive. However, like other related image generation methods, it could still potentially be misused in the realm of human impersonation. A notorious example are so-called "deep fakes" that have been used, for example, to create pornographic "undressing" applications. We strongly disapprove of any actions aimed at producing deceptive or harmful content featuring real individuals. Besides, generative methods have the capacity to be harnessed for other malicious intentions, including harassment and misinformation spread [20], and give rise to significant concerns pertaining to societal and cultural exclusion as well as biases [83, 82]. These considerations guide our decision not to release the source code or a public demo at this point in time.

Furthermore, the immediate availability of mass-produced high-quality images can be used to spread misinformation and spam, which in turn can be used for targeted manipulation in social media. Datasets are crucial for deep learning as they are the main input of information [101, 92, 93, 97]. Large-scale data requirements of text-to-image models have led researchers to rely heavily on large, mostly uncurated, web-scraped datasets. While this approach has enabled rapid algorithmic advances recently, datasets of this nature have been critiqued and contested along various ethical dimensions. One should consider the ability to curate the database to exclude (or explicitly contain) potential harmful source images. Creating a public API could offer a cheaper way to offer a safe model than retraining a model on a filtered subset of the training data or doing difficult prompt engineering. Conversely, including only harmful content is an easy way to build a toxic model.

A.2 Guidance Scale *vs*. FID

To further demonstrate the effectiveness of our proposed context prediction, we quantitatively conduct evaluations about the trade-off between MS-COCO zero-shot FID [26] and CLIP scores. The results in Figure 6 indicate that the guidance hurts the diversity of GLIDE much more than DALL-E 2 and ConPreDiff. The phenomenon reveals that the proposed ConPreDiff can overall improve the generation quality of diffusion models.

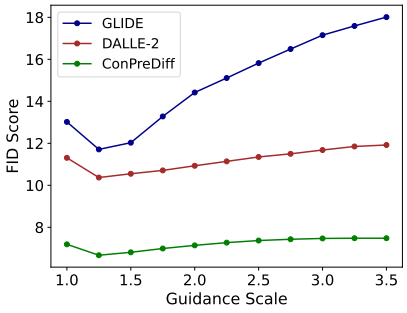

Figure 6: Trade-off between guidance scale and FID.

A.3 More Quantitative Results

We list the unconditional generation results on FFHQ, CelebA-HQ, LSUN-Churches, and LSUN-Bedrooms in Tab. 3. We find CONPREDIFF consistently outperforms previous methods, demonstrating the effectiveness of the CONPREDIFF.

Table 3: Evaluation results for unconditional image synthesis.

FFHQ 256×256			
Method	FID ↓	Prec. ↑	Recall ↑
ImageBART[16]	9.57	-	-
U-Net GAN (+aug) [72]	7.6	-	-
UDM [39]	5.54	-	-
StyleGAN [36]	4.16	0.71	0.46
ProjectedGAN [71]	3.08	0.65	0.46
LDM [65]	4.98	0.73	0.50
CONPREDIFF	**2.24**	**0.81**	**0.61**
LSUN-Bedrooms 256×256			
Method	FID ↓	Prec. ↑	Recall ↑
ImageBART [16]	5.51	-	-
DDPM [28]	4.9	-	-
UDM [39]	4.57	-	-
StyleGAN [36]	2.35	0.59	0.48
ADM [14]	1.90	0.66	0.51
ProjectedGAN [71]	1.52	0.61	0.34
LDM-4 [65]	2.95	0.66	0.48
CONPREDIFF	**1.12**	**0.73**	**0.59**
CelebA-HQ 256×256			
Method	FID ↓	Prec. ↑	Recall ↑
DC-VAE [55]	15.8	-	-
VQGAN+T. [17] (k=400)	10.2	-	-
PGGAN [43]	8.0	-	-
LSGM [87]	7.22	-	-
UDM [39]	7.16	-	-
LDM [65]	5.11	0.72	0.49
CONPREDIFF	**3.22**	**0.83**	**0.57**
LSUN-Churches 256×256			
Method	FID ↓	Prec. ↑	Recall ↑
DDPM [28]	7.89	-	-
ImageBART [16]	7.32	-	-
PGGAN [43]	6.42	-	-
StyleGAN [36]	4.21	-	-
StyleGAN2 [37]	3.86	-	-
ProjectedGAN [71]	**1.59**	0.61	0.44
LDM [65]	4.02	0.64	0.52
CONPREDIFF	**1.78**	**0.74**	**0.61**

A.4 More Synthesis Results

We visualize more text-to-image synthesis results on MS-COCO dataset in Figure 7. We observe that compared with previous powerful LDM and DALL-E 2, our CONPREDIFF generates more natural and smooth images that preserve local continuity.

Figure 7: Synthesis examples demonstrating text-to-image capabilities of for various text prompts.

A.5 Human Evaluations

As demonstrated in qualitative results, our CONPREDIFF is able to synthesize realistic diverse, context-coherent images. However, using FID to estimate the sample quality is not always consistent with human judgment. Therefore, we follow the protocol of previous works [104, 68, 62], and conduct systematic human evaluations to better assess the generation capacities of our CONPREDIFF from the aspects of image photorealism and image-text alignment. We conduct side-by-side human evaluations, in which well-trained users are presented with two generated images for the same prompt and need to choose which image is of higher quality and more realistic (image photorealism) and which image better matches the input prompt (image-text alignment). For evaluating the coherence of local context, we propose a new evaluation protocol, in which users are presented with 1000 pairs of images and must choose which image better preserves local pixel/semantic continuity. The evaluation results are in Tab. 4, CONPREDIFF performs better in pairwise comparisons against both Improved VQ-Diffusion and Imagen. We find that CONPREDIFF is preferred in terms of all three evaluations, and CONPREDIFF is strongly preferred regarding context coherence, demonstrating that preserving local neighborhood context advances sample quality and semantic alignment.

Table 4: Human evaluation comparing CONPREDIFF to Improved VQ-Diffusion and Imagen.

	Improved VQ-Diffusion	Imagen
Image Photorealism	72%	65%
Image-Text Alignment	68%	63%
Context Coherence	84%	78%