Appendix

G Ablation

G.1 Comparison to DiT and LDM

On contrary to LDM [59] and DiT [56], the latent DiffiT does not rely on shift and scale, as in AdaLN [56], or concatenation to incorporate time embedding into the denoising networks. However, DiffiT uses a time-dependent self-attention (*i.e.* TMSA) to jointly learn the spatial and temporal dependencies. In addition, DiffiT proposes both image and latent space models for different image generation tasks with different resolutions with SOTA performance. Specifically, as shown in Table S.1, DiffiT significantly outperforms LDM [59] and DiT [56] by 31.26% and 51.94% in terms of FID score on ImageNet-256 [15] dataset. In addition, DiffiT outperforms DiT [56] by 13.85% on ImageNet-512 [15] dataset. Hence, these benchmarks validate the effectiveness of the proposes architecture and TMSA design in DiffiT model as opposed to previous SOTA for both CNN and Transformer-based diffusion models.

Model	Class	ImageNet-256				ImageNet-512			
		FID ↓	IS ↑	Precision ↑	Recall ↑	FID ↓	IS ↑	Precision ↑	Recall ↑
LDM-4-G [59]	Diffusion	3.60	247.67	0.87	0.48	-	-	-	-
DiT-XL/2-G [56]	Diffusion	2.27	278.24	0.83	0.57	3.04	240.82	0.84	0.54
DiffiT	Diffusion	**1.73**	276.49	0.80	0.62	**2.67**	252.12	0.83	0.55

Table S.1 – Comparison of image generation performance against state-of-the-art models on ImageNet-256 and ImageNet-512 dataset. The latent DiffiT model achieves SOTA performance in terms of FID score on ImageNet-256 dataset.

H Architecture

H.1 Image Space

We provide the details of blocks and their corresponding output sizes for both the encoder and decoder of the DiffiT model in Table S.2 and Table S.3, respectively. The presented architecture details denote models that are trained with 64×64 resolution. Without loss of generality, the architecture can be extended for 32×32 resolution. For FFHQ-64 [36] dataset, the values of L_1, L_2, L_3 and L_4 are 4, 4, 4, and 4 respectively. For CIFAR-10 [45] dataset, the architecture spans across three different resolution levels (*i.e.* 32, 16, 8), and the values of L_1, L_2, L_3 are 4, 4, 4 respectively. Please refer to the paper for more information regarding the architecture details.

Table S.2 – Detailed description of components in DiffiT encoder for models that are trained at 64×64 resolution.

Component Description	Output size
Input	$64 \times 64 \times 3$
Tokenizer	$64 \times 64 \times 128$
DiffiT ResBlock $\times L_1$	$64 \times 64 \times 128$
Downsampler	$32 \times 32 \times 128$
DiffiT ResBlock $\times L_2$	$32 \times 32 \times 256$
Downsampler	$16 \times 16 \times 128$
DiffiT ResBlock $\times L_3$	$16 \times 16 \times 256$
Downsampler	$8 \times 8 \times 256$
DiffiT ResBlock $\times L_4$	$8 \times 8 \times 256$

Table S.3 – Detailed description of components in DiffiT decoder for models that are trained at 64×64 resolution.

Component Description	Output size
Input	$8 \times 8 \times 256$
Upsampler	$16 \times 16 \times 256$
DiffiT ResBlock $\times L_3$	$16 \times 16 \times 256$
Upsampler	$32 \times 32 \times 256$
DiffiT ResBlock $\times L_2$	$32 \times 32 \times 256$
Upsampler	$64 \times 64 \times 256$
DiffiT ResBlock $\times L_1$	$64 \times 64 \times 128$
Head	$64 \times 64 \times 3$

H.2 Latent Space

In Fig S.1, we illustrate the architecture of the latent DiffiT model. Our model is comparable to DiT-XL/2-G variant which 032 uses a patch size of 2. Specifically, we use a depth of 30 layers with hidden size dimension of 1152, number of heads dimension of 16 and MLP ratio of 4. In addition, for the classifier-free guidance implementation, we only apply the guidance to the first three input channels with a scale of $(1 + \mathbf{x})$ where $\mathbf{x}$ is the input latent.

I Implementation Details

I.1 Image Space

We strictly followed the training configurations and data augmentation strategies of the EDM [34] model for the experiments on CIFAR10 [45], and FFHQ-64 [36] datasets, all in an unconditional setting. All the experiments were trained for 200000 iterations with Adam optimizer [41] and used PyTorch framework and 8 NVIDIA A100 GPUs. We used batch sizes of 512 and 256, learning rates of

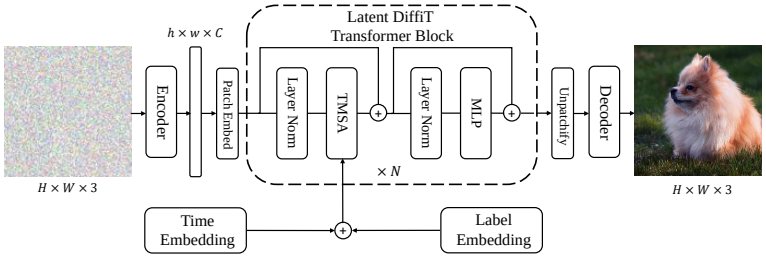

Fig. S.1 – Overview of the latent DiffiT framework.

1×10^{-3} and 2×10^{-4} and training images of sizes 32×32 and 64×64 on experiments for CIFAR10 [45] and FFHQ-64 [36] datasets, respectively.

We use the deterministic sampler of EDM [34] model with 18, 40 and 40 steps for CIFAR-10 and FFHQ-64 datasets, respectively. For FFHQ-64 dataset, our DiffiT network spans across 4 different stages with 1, 2, 2, 2 blocks at each stage. We also use window-based attention TMSA with local window size of 8 at each stage. For CIFAR-10 dataset, the DiffiT network has 3 stages with 2 blocks at each stage. Similarly, we compute attentions on local windows with size 4 at each stage. Note that for all networks, the resolution is decreased by a factor of 2 in between stages. However, except for when transitioning from the first to second stage, we keep the number of channels constant in the rest of the stages to maintain both the number of parameters and latency in our network. Furthermore, we employ traditional convolutional-based downsampling and upsampling layers for transitioning into lower or higher resolutions. We achieved similar image generation performance by using bilinear interpolation for feature resizing instead of convolution. For fair comparison, in all of our experiments, we used the FID score which is computed on 50K samples and using the training set as the reference set.

I.2 Latent Space

We employ learning rates of 3×10^{-4} and 1×10^{-4} and batch sizes of 256 and 512 for ImageNet-256 and ImageNet-512 experiments, respectively. We also use the exponential moving average (EMA) of weights using a decay of 0.9999 for both experiments. We also use the same diffusion hyper-parameters as in the ADM [16] model. For a fair comparison, we use the DDPM [28] sampler with 250 steps and report FID-50K for both ImageNet-256 and ImageNet-512 experiments.

J Qualitative Results

We illustrate visualization of generated images for CIFAR-10 [45] and FFHQ-64 [36] datasets in Figures S.2 and S.3, respectively. In addition, in Figures S.4, S.5, S.6 and S.7, we visualize the the generated images by the latent DiffiT model

for ImageNet-512 [15] dataset. Similarly, the generated images for ImageNet-256 [15] are shown in Figures S.8, S.9 and S.10. We observe that the proposed DiffiT model is capable of capturing fine-grained details and produce high fidelity images across these datasets.

Fig. S.2 – Visualization of uncurated generated images for CIFAR-10 [45] dataset. Best viewed in color.

Fig. S.3 – Visualization of uncurated generated images for FFHQ-64 [36] dataset. Best viewed in color.

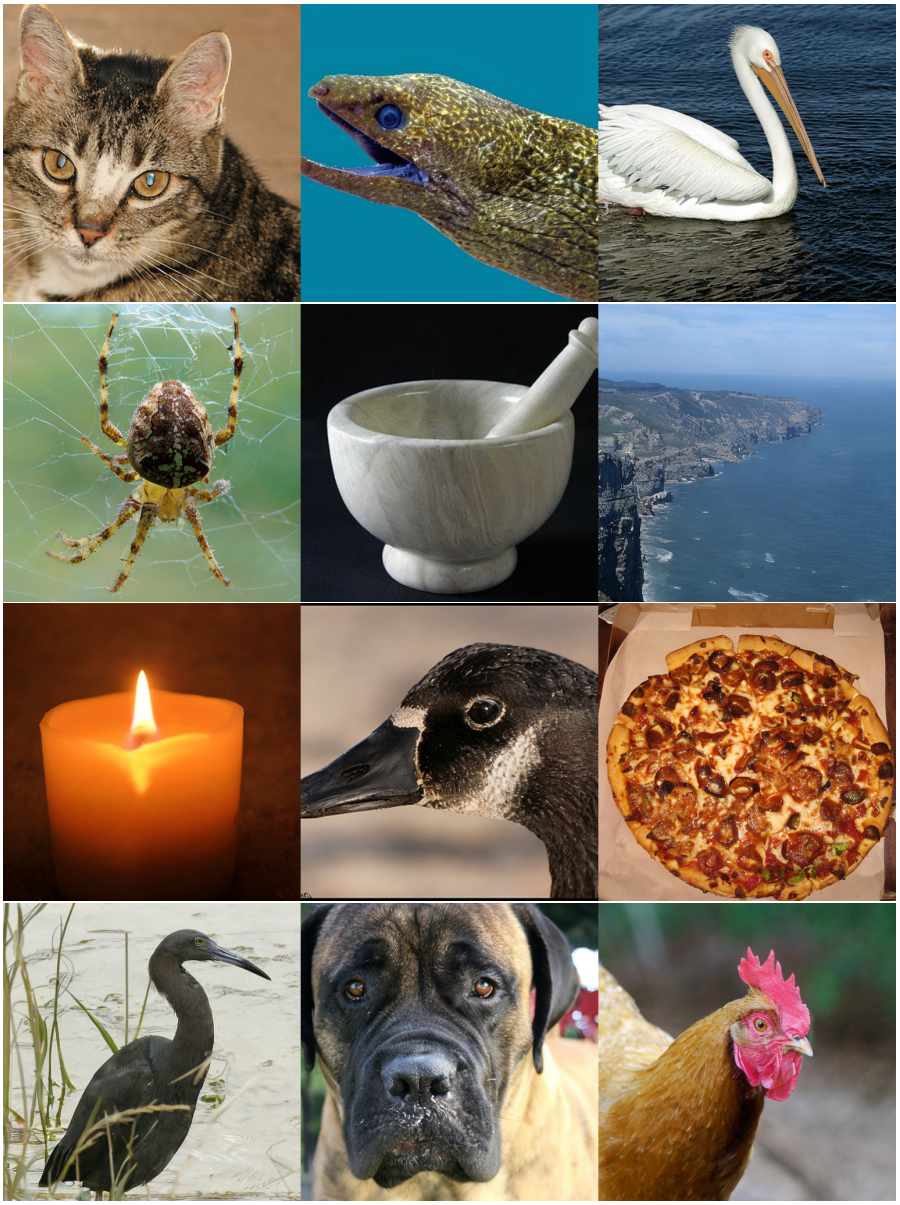

Fig. S.4 – Visualization of uncurated generated 512×512 images on ImageNet [15] dataset by latent DiffiT model. Images are randomly sampled. Best viewed in color.

Fig. S.5 – Visualization of uncurated generated 512×512 images on ImageNet [15] dataset by latent DiffiT model. Images are randomly sampled. Best viewed in color.

Fig. S.6 – Visualization of uncurated generated 512×512 images on ImageNet [15] dataset by latent DiffiT model. Images are randomly sampled. Best viewed in color.

Fig. S.7 – Visualization of uncurated generated 512×512 images on ImageNet [15] dataset by latent DiffiT model. Images are randomly sampled. Best viewed in color.

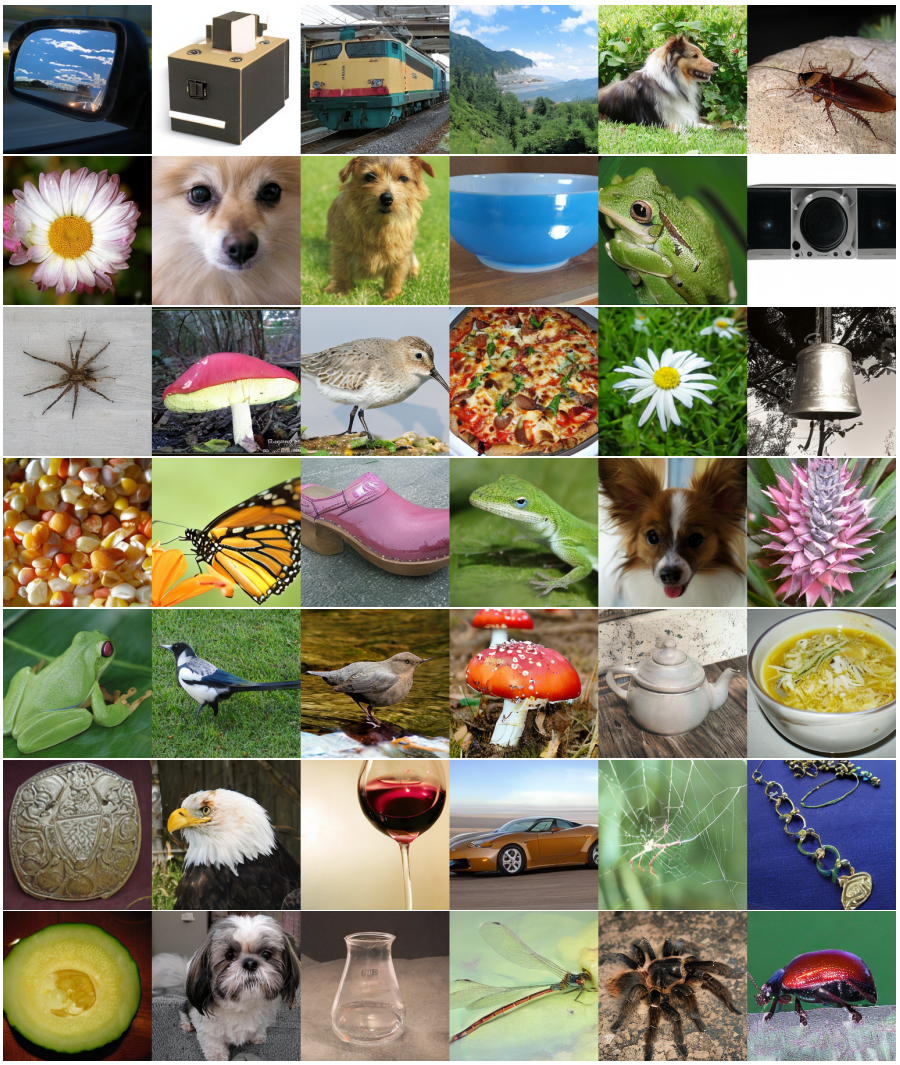

Fig. S.8 – Visualization of uncurated generated 256×256 images on ImageNet [15] dataset by latent DiffiT model. Images are randomly sampled. Best viewed in color.

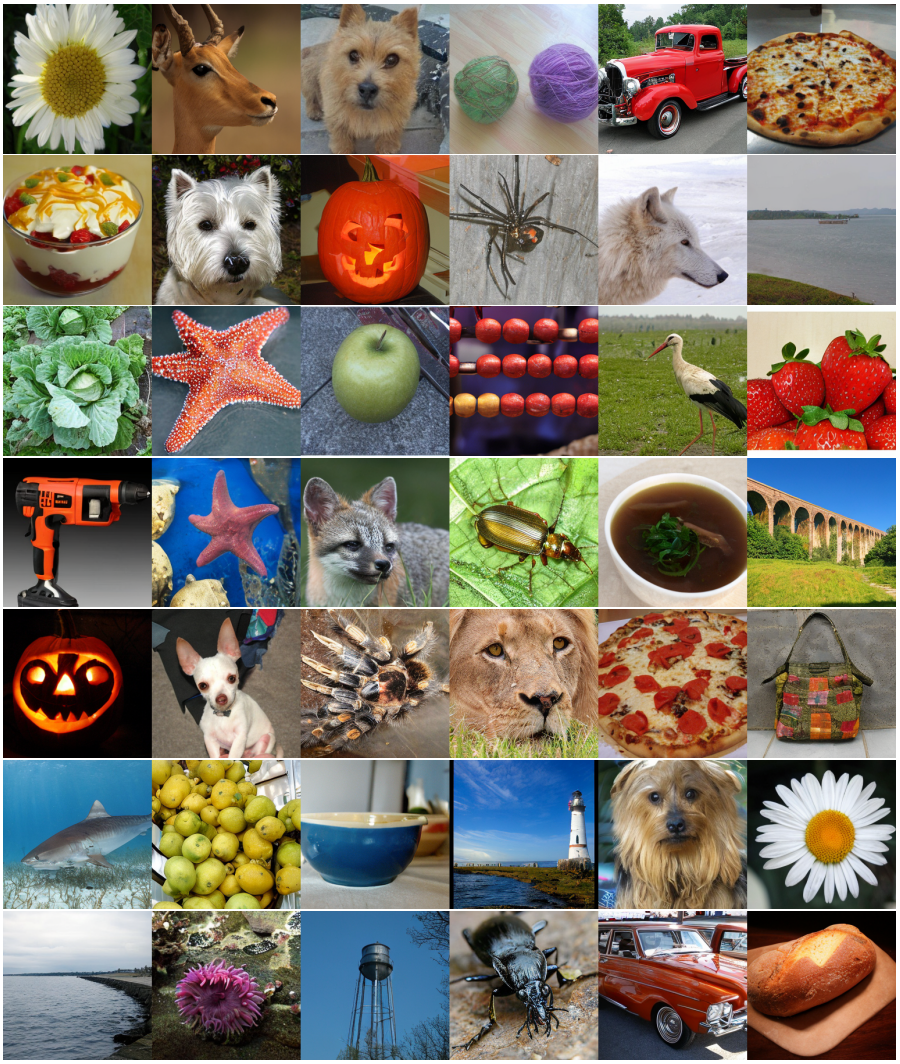

Fig. S.9 – Visualization of uncurated generated 256×256 images on ImageNet [15] dataset by latent DiffiT model. Images are randomly sampled. Best viewed in color.

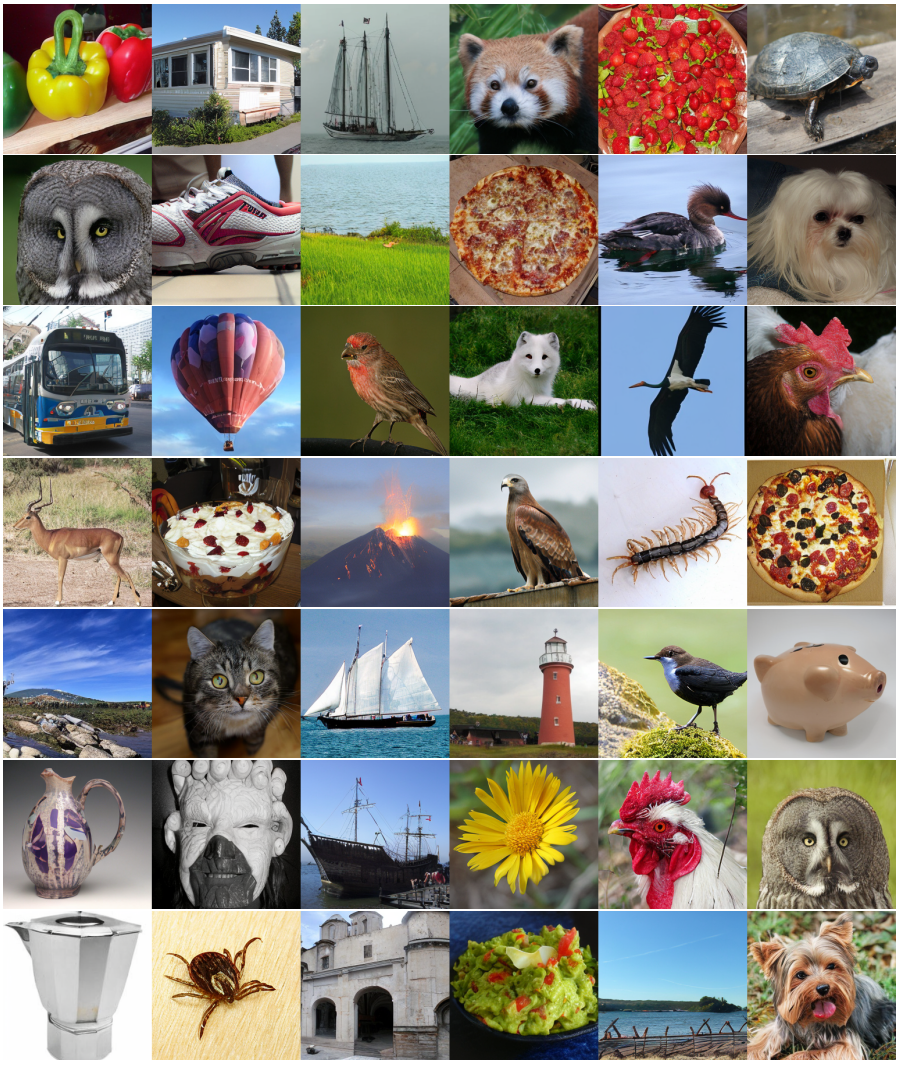

Fig. S.10 – Visualization of uncurated generated 256×256 images on ImageNet [15] dataset by latent DiffiT model. Images are randomly sampled. Best viewed in color.