SUPPLEMENTARY MATERIAL: PATCHED DENOISING DIFFUSION MODELS FOR HIGH-RESOLUTION IMAGE SYNTHESIS

Anonymous authors
Paper under double-blind review

A MODEL HYPERPARAMETERS

For model architecture, we base our diffusion model from Dhariwal & Nichol (2021) with changes of taking global semantic condition and positional embedding. The hyperparameters for the main denoising U-Net model are specified in Table 1. Since the model is resolution agnostic, the main architectures for all datasets keep the same. We adopt two methods to obtain global semantic conditions: for relatively low-resolution images, an encoder is trained with architecture borrowed from the first half of the U-Net model, and the architecture details of the encoder are shown in Table 2. For high-resolution images such as 1024×512, a pretrained image encoder is used to avoid scaling up the overall model size. We use ViT-B/16 in CLIP to obtain the image embeddings and optimize them during training. For position embeddings, we use sinusoidal positional embeddings. Time embedding and positional embedding are concatenated and modulated into ResBlocks together with the global code.

For realizing unconditional image synthesis, a latent diffusion model is trained on semantic embeddings. The implementation is based on the one proposed in Preechakul et al. (2022) with MLP + skip connections architecture. The parameter details are specified in Table 3.

B MORE QUALITATIVE RESULTS

We provide more unconditional sampling results with the models trained on our self-collected nature images (1024×512) in Figure 1-3. We also provide results on LHQ(1024×1024) and FFHQ(1024×1024) in Figures 5 and 4 respectively as well as three other standard benchmarks with a resolution of 256×256: LSUN-Bedroom (Figure 6), LSUN-Church (Figure 7), and FFHQ (Figure 8).

Parameter	Patch-DM
Patch input size	3×64×64
Channel multiplier	[1, 2, 4, 8]
Net channel	64
ResBlock number	2
Attention resolution	16
Batch size	16
Diffusion steps	1000
Noise scheduler	Linear
Learning rate	0.0001
Optimizer	Adam

Table 1: Model Architecture for diffusion model.

Parameter	Semantic Encoder
Input size	3×256×256
Channel multiplier	[1, 2, 4, 8, 8]
Net channel	64
ResBlock number	2
Attention resolution	16
Global condition dimension	512
Batch size	16
Learning rate	0.0001
Optimizer	Adam

Table 2: Model Architecture for image semantic encoder.

Parameter	Latent Diffusion
Input size	512
MLP layers	10
MLP hidden size	2048
Noise scheduler	Constant 0.008
Batch size	256
Learning rate	0.0001
Optimizer	Adam (weight decay 0.01)

Table 3: Model Architecture for latent diffusion model.

Figure 1: Additional qualitative results on self-collected nature dataset (1024×512).

Figure 2: Additional qualitative results on self-collected nature dataset (1024×512).

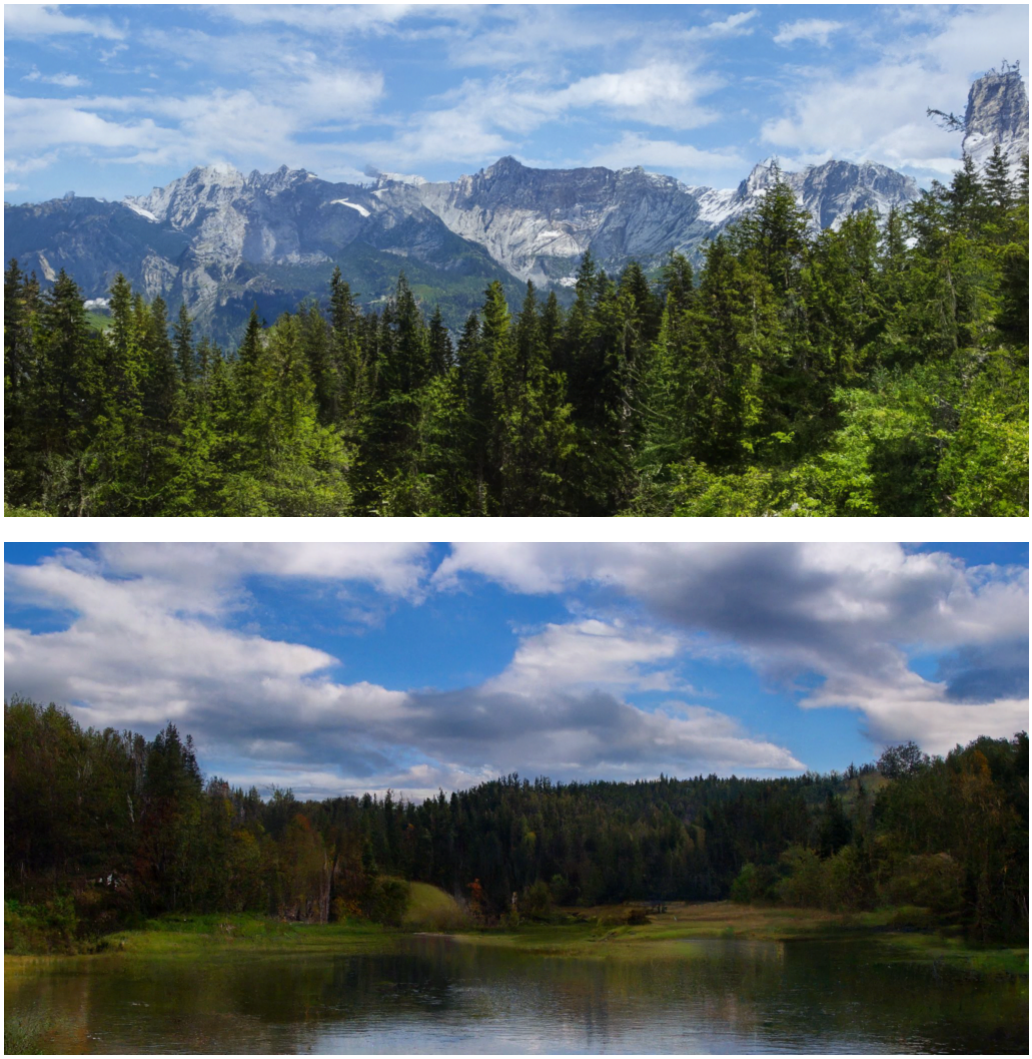

Figure 3: Additional qualitative results on self-collected nature dataset (1024×512).

Figure 4: Additional qualitative results on high resolution FFHQ dataset (1024×1024).

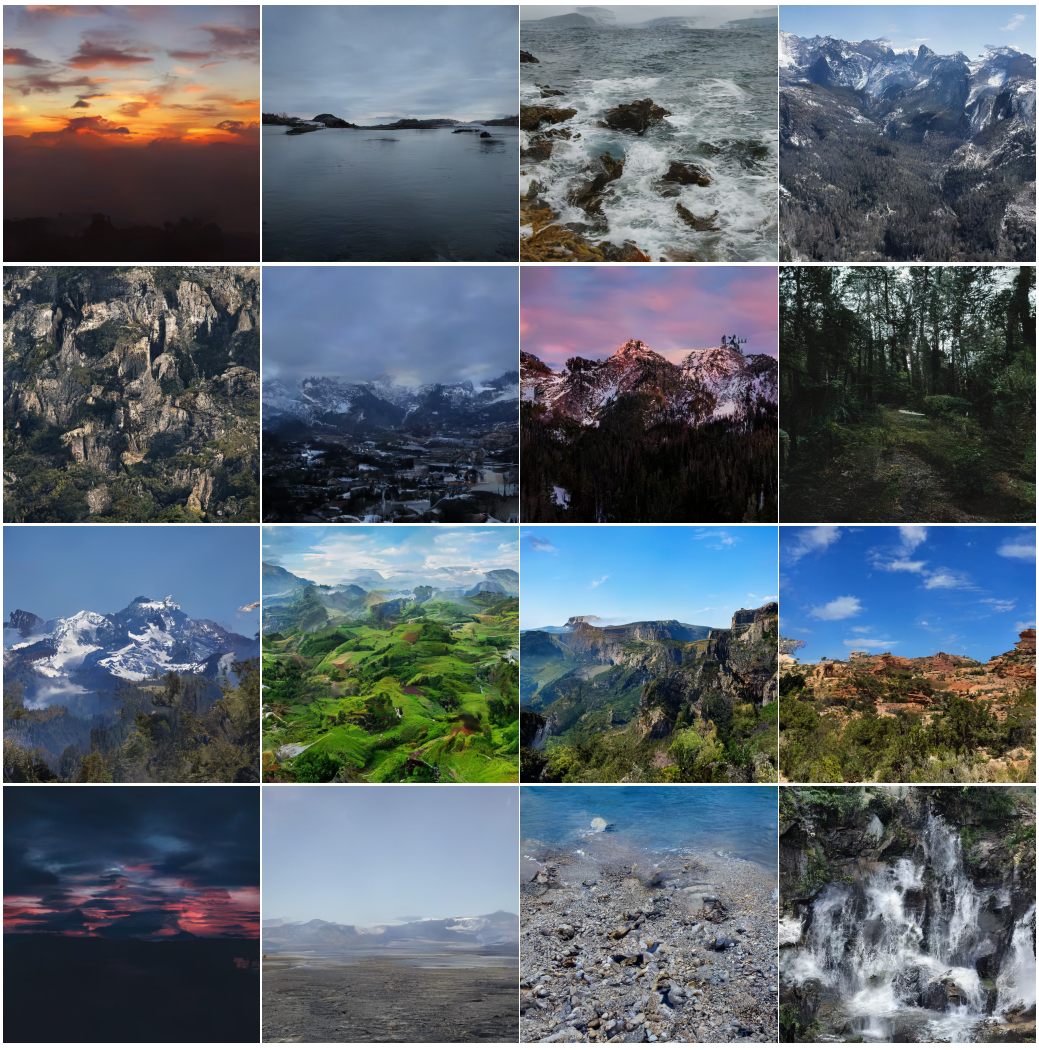

Figure 5: Additional qualitative results on high resolution LHQ dataset (1024×1024).

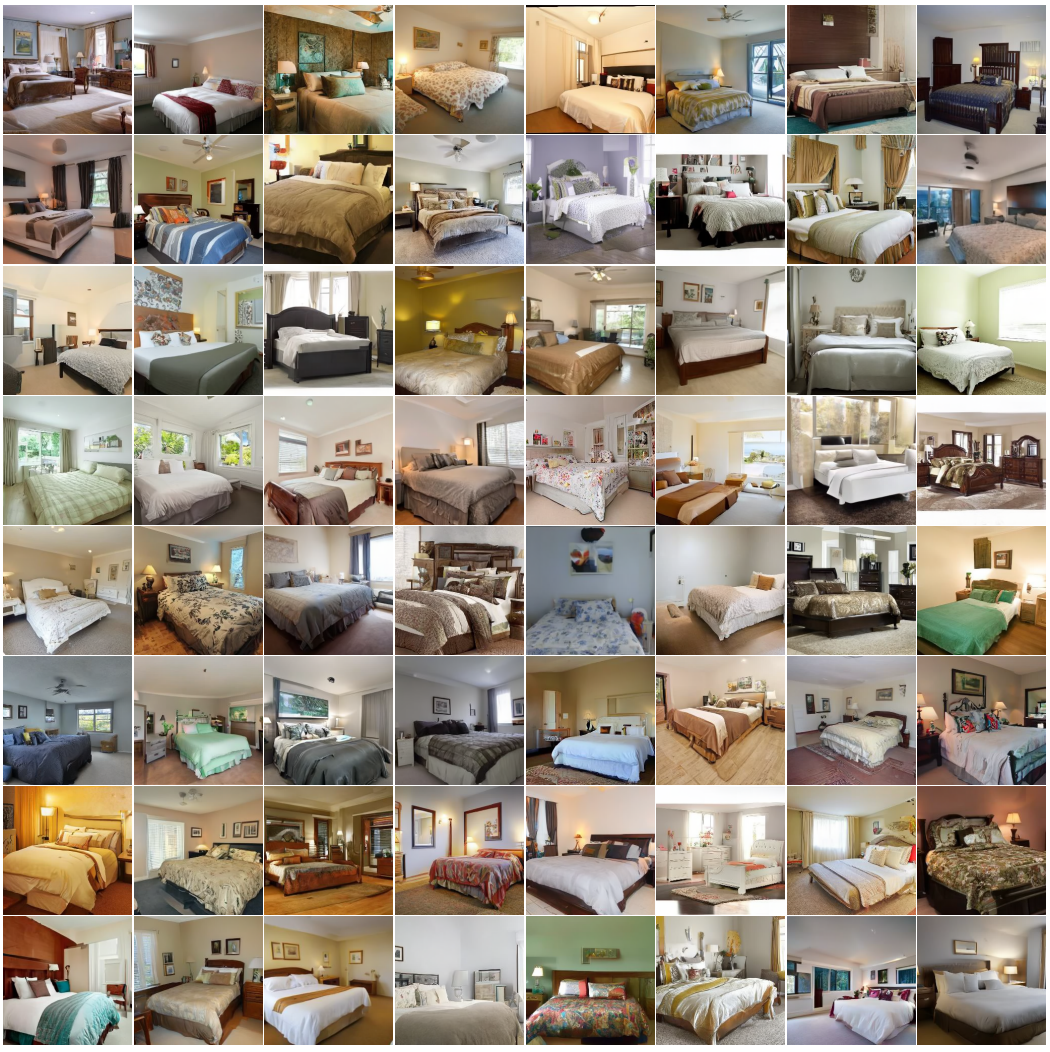

Figure 6: Additional qualitative results on LSUN-Bedroom (256×256).

Figure 7: Additional qualitative results on LSUN-Church (256×256).

Figure 8: Additional qualitative results on FFHQ (256×256).

C FAILED CASES COMPARISON ON FFHQ1024

To provide a more thorough understanding of the limitations of our proposed method, we compare our FFHQ1024 results with other non-patch-based methods. We use FFHQ1024 as it's a more structural dataset while other landscape datasets are not that structural, thus FFHQ1024 could better show how well the models learn the structural information. As there lacks results for FFHQ1024 using diffusion models, we compare our results with two state-of-the-art GAN-based methods on FFHQ1024: StyleGAN3(Karras et al., 2021) and StyleGAN-XL(Sauer et al., 2022) in Figure 9. It can be seen that non-patch GAN-based methods generally perform well on global consistency while ours may not perform well in such cases. For example, in the first row of our results, the wrinkles around the mouth are asymmetrical. And in the second row of our results, the eyes/eyeglasses are asymmetrical. As our method uses a patch size of 64×64 and there are a total of $16\times16=256$ patches, the global consistency sometimes may not perform well. We think one can further improve this by utilizing a larger patch size or introducing better global-consistency-enforcing mechanisms which we leave for future work.

StyleGAN3-T StyleGAN-XL Our Failed Cases

Figure 9: Comparison on FFHQ1024 with StyleGAN3-T, StyleGAN-XL and our failed cases. Ours might fail on global consistency such as the wrinkles around the mouth in the first row and the eyes/glasses in the second row.

REFERENCES

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. NeurIPS*, 2021.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10619–10629, 2022.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022.