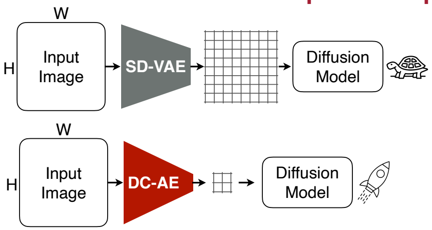
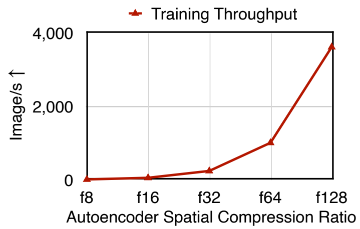
Deep Compression Autoencoder for Efficient High-Resolution Diffusion Models

Junyu Chen[1,2*], Han Cai[3*†], Junsong Chen[3], Enze Xie[3], Shang Yang[1], Haotian Tang[1], Muyang Li[1], Song Han[1,3]
[1]Massachusetts Institute of Technology, [2]Tsinghua University, [3]NVIDIA

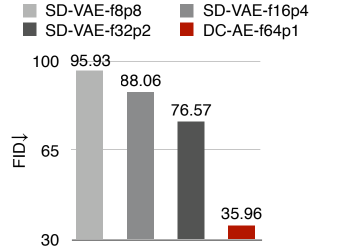
DC-AE: Latent Space Compression for Accelerating Latent Diffusion Models



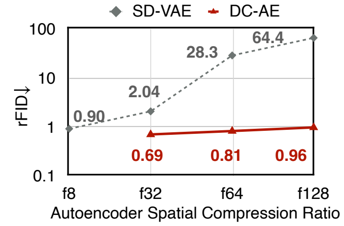
- Previous diffusion model acceleration techniques focus on compressing the diffusion models.
- DC-AE introduces a new direction for diffusion model acceleration: latent space compression.



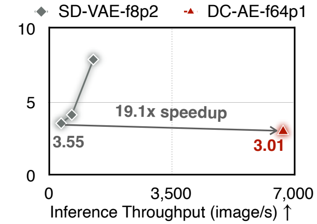
- Increasing the autoencoder's spatial compression ratio effectively improves diffusion models' training efficiency by producing a latent space with fewer tokens.



- Shifting the spatial compression from the diffusion model to the autoencoder leads to better image generation quality.



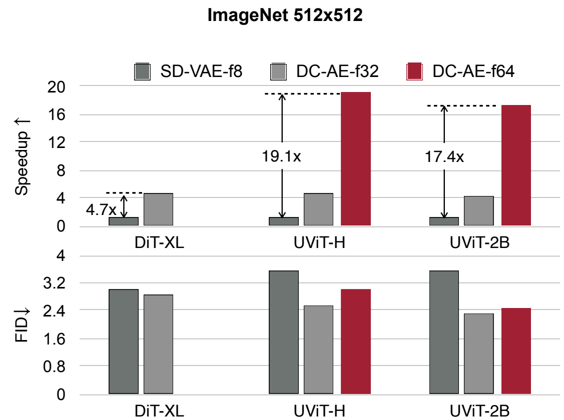
- Challenge: Accuracy Degradation of High Spatial-Compression Autoencoders.



- DC-AE delivers a significant speedup by compressing the latent space.

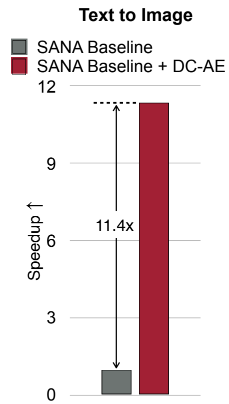
Better Reconstruction Performance

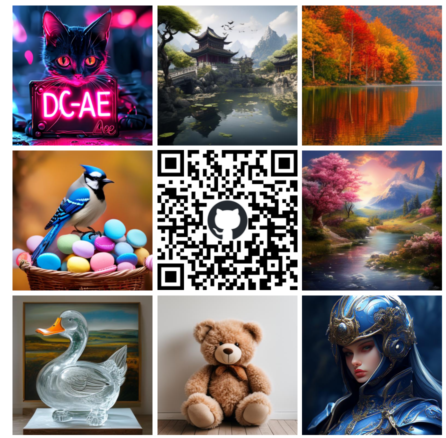
- DC-AE-f64 provides significantly better reconstruction quality than SD-VAE-f64.



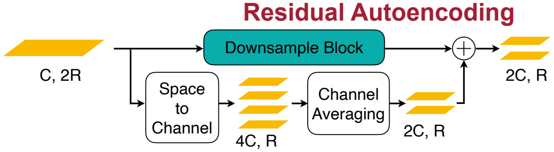
- DC-AE provides up to 19.1x speedup while achieving a better FID.

Latent Diffusion Generation Speedup

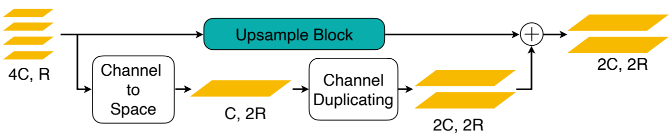




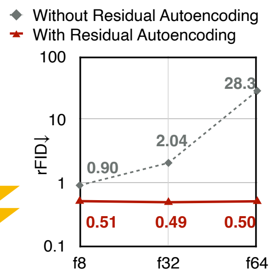
Residual Autoencoding



(a) DC-AE Downsample Block with Residual Autoencoding

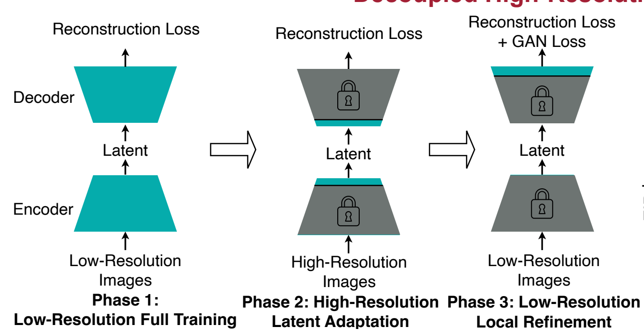


(b) DC-AE Upsample Block with Residual Autoencoding

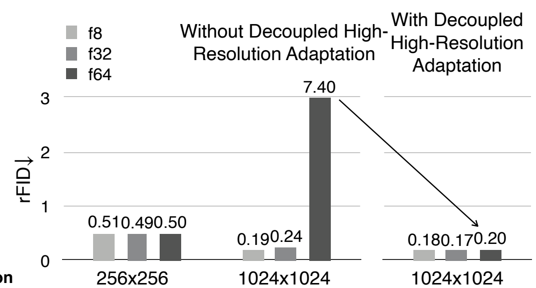


- DC-AE introduces residual autoencoding to address degradation of reconstruction accuracy. It lets the downsample (or upsample) blocks learn residuals based on space-to-channel (or channel-to-space).

Decoupled High-Resolution Adaptation



- A three-stage training pipeline to efficiently fine-tune high spatial-compression autoencoders to adapt to high-resolution images.



- It effectively reduces the generalization gap from low-resolution to high-resolution.