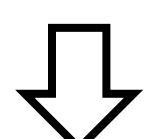


Introduction

- Diffusion Models (DMs)^[1] can generate high-quality and diverse images
- Pixel-based DMs add or remove noises over **pixel values**

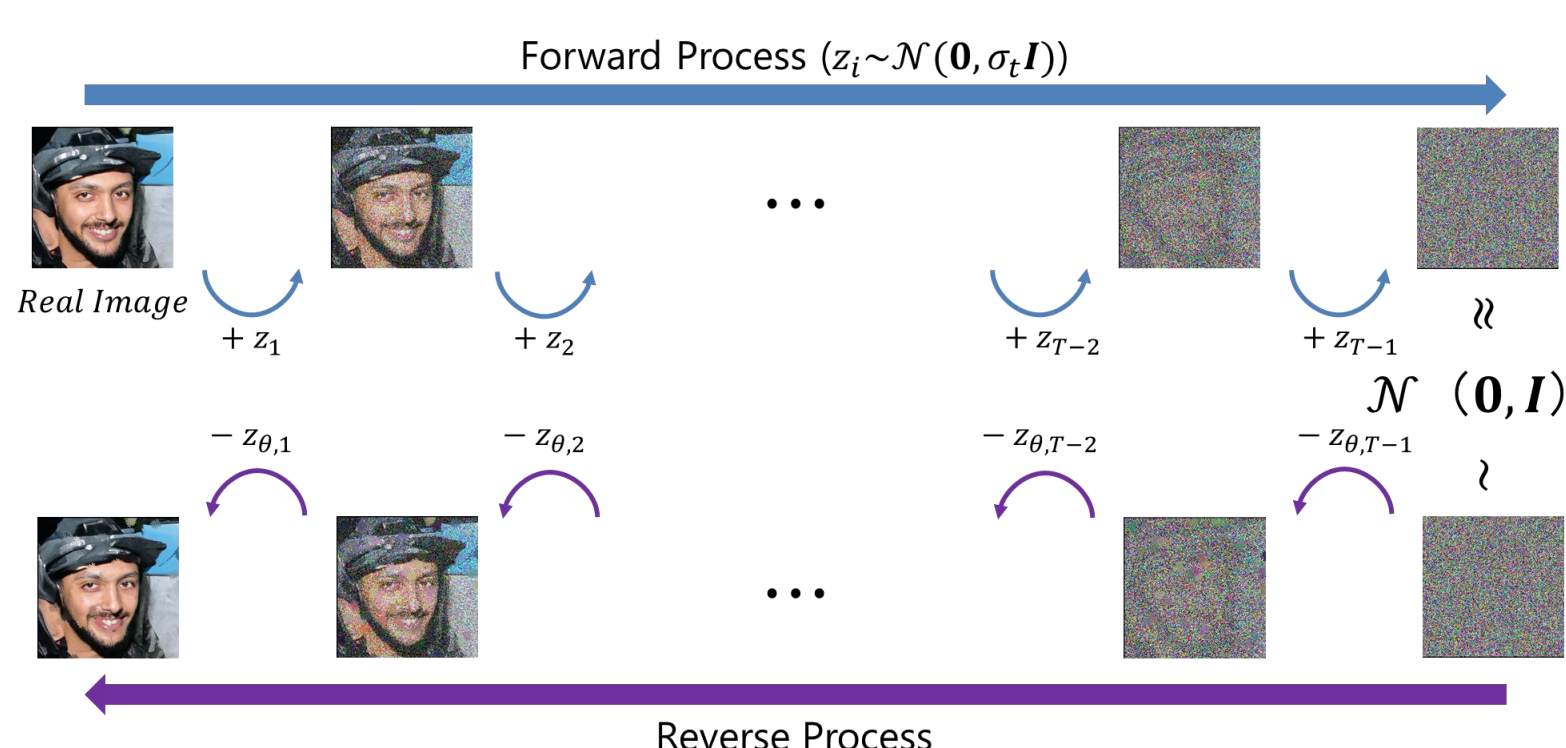


Requires large memory consumption to generate high-resolution images

We tackle this problem by “**patch-by-patch generation**” using spatially smaller patches

Key features are ...

- **Positional information**
- **Content information**



Generated images by DDPM^[1]

Related Works

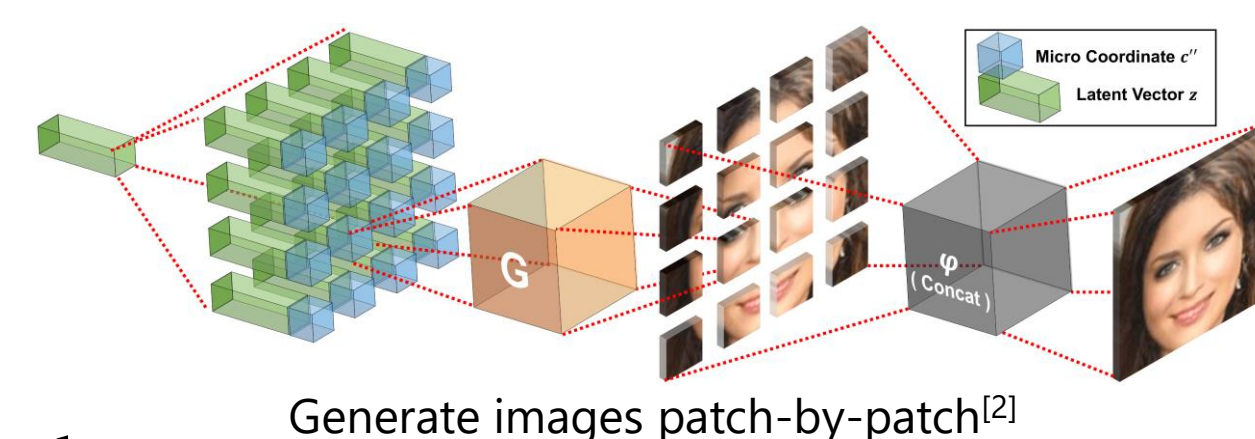
Patch-based Approaches

- COCO-GAN^[2]
 - Fully patch-by-patch approach in GANs
 - Generates awesome quality images
 - **Memory efficient and Parallelizable**

➔ **Enabled by the supervision of the Discriminator**

- Patch Diffusion^[3]
 - Trains diffusion models with patches cropped at random positions
 - The network is conditioned by **coordinates**
 - Fast training and convergence

➔ **This method is effective for training**

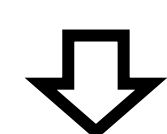


Generate images patch-by-patch^[2]

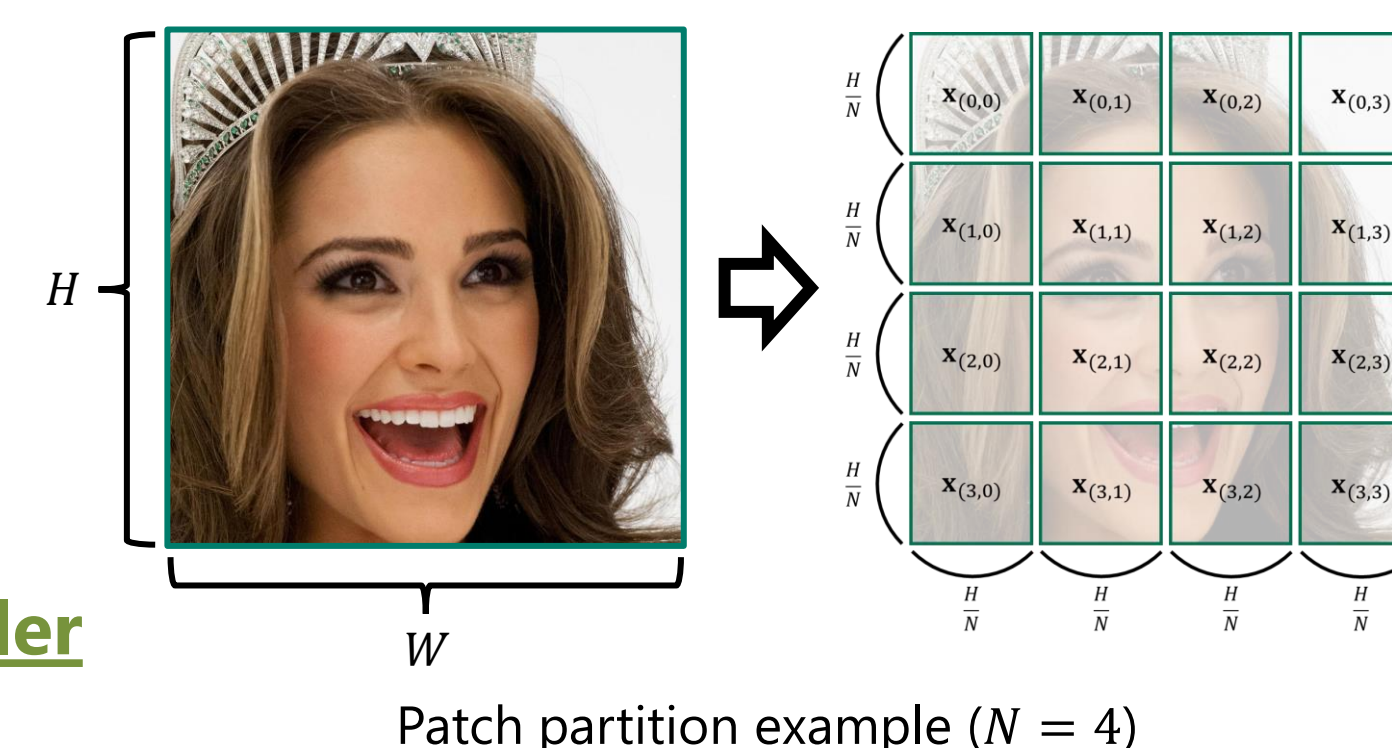
Methods

Patch Partition

- Divide the entire image into patches
- Grid-like manner and non-overlap
- Training and inference are operated only on patches



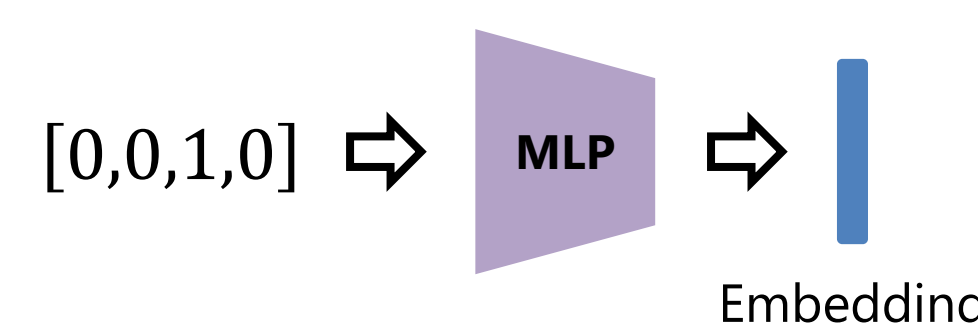
Making tensors passed in the network smaller



Patch partition example ($N = 4$)

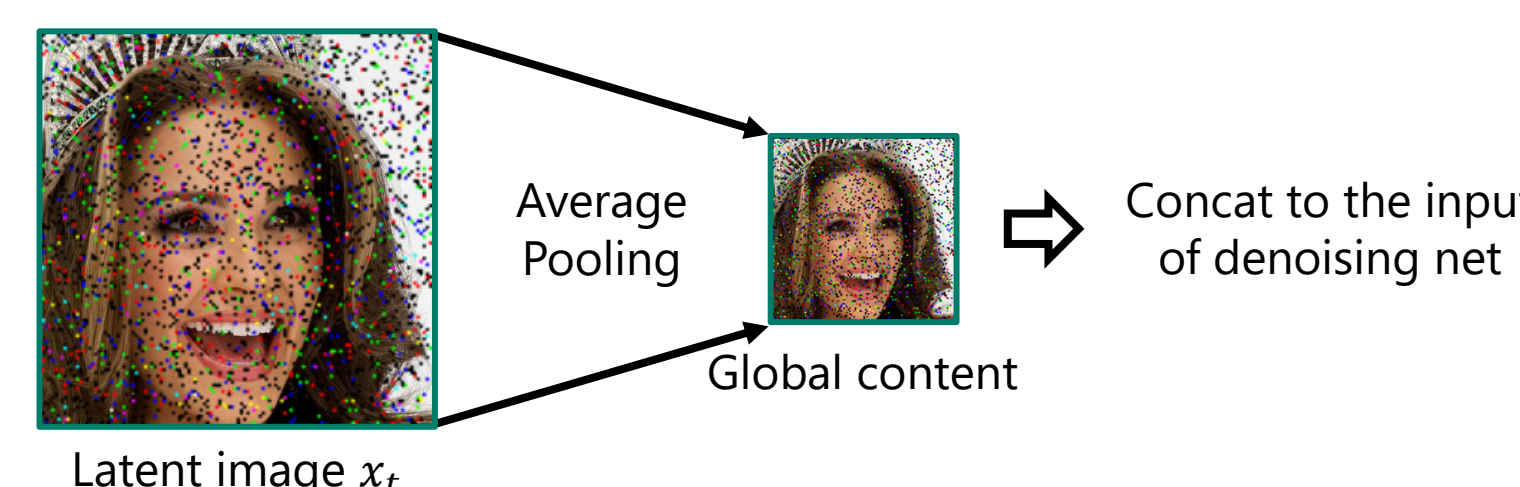
Positional Condition

- Condition the **patch location** to the denoising network
- Define unique index $s = i \times N + j$
- Convert s to a one-hot vector and pass it to MLP

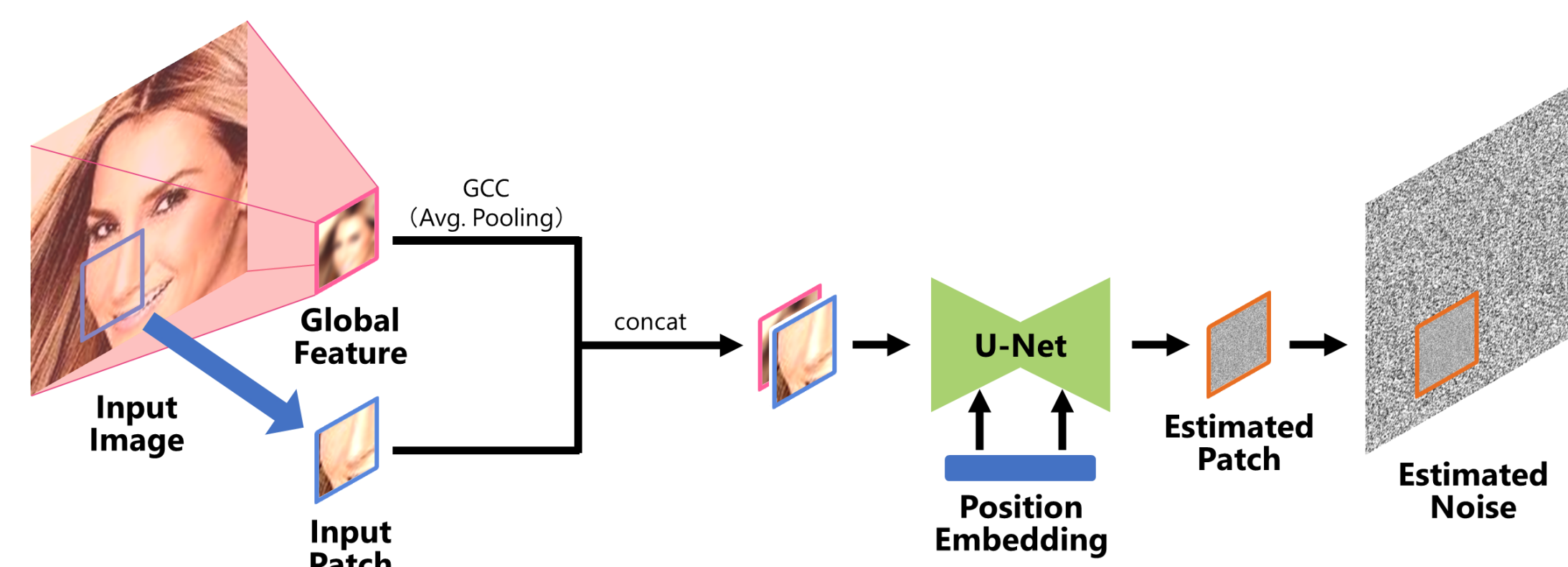


Global Content Conditioning (GCC)

- Share “**content**” across patches
- Apply avg. pooling to the entire image to obtain a feature map
- Concatenate the feature map with the cropped patches



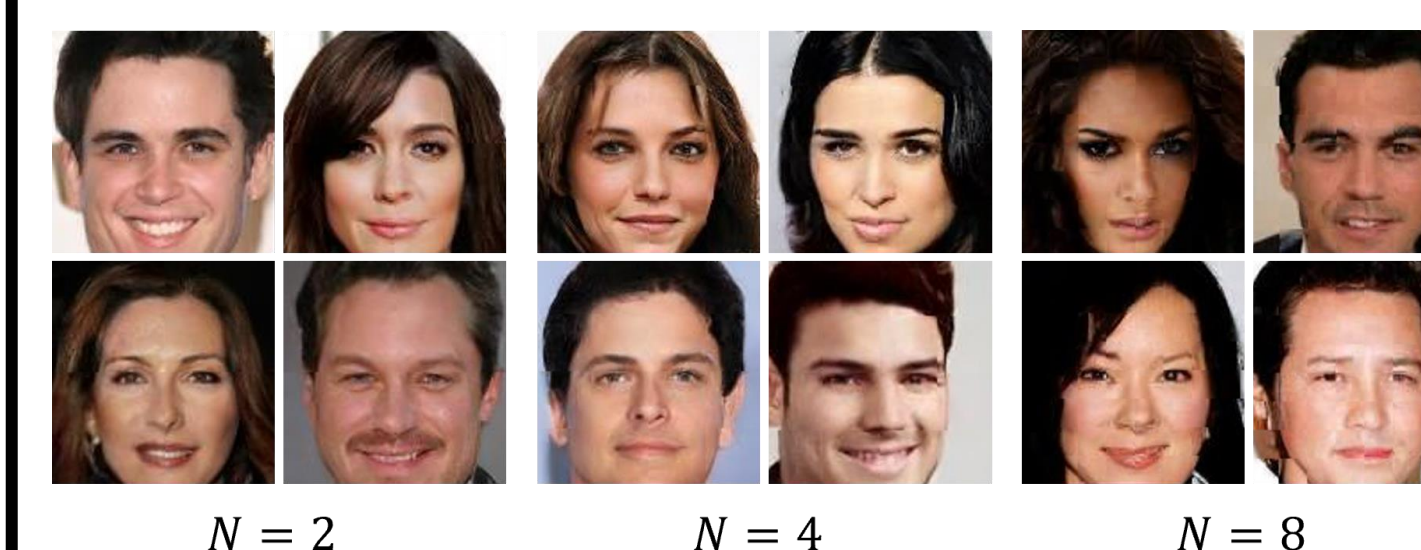
Overall proposed method



- **Patch-wise network evaluation** with positional and global content conditions
- Latent image x_t is split into patches, then passed into U-Net
- The patch-wise results are concatenated to obtain an entire noise estimation

Results

CelebA^[4]



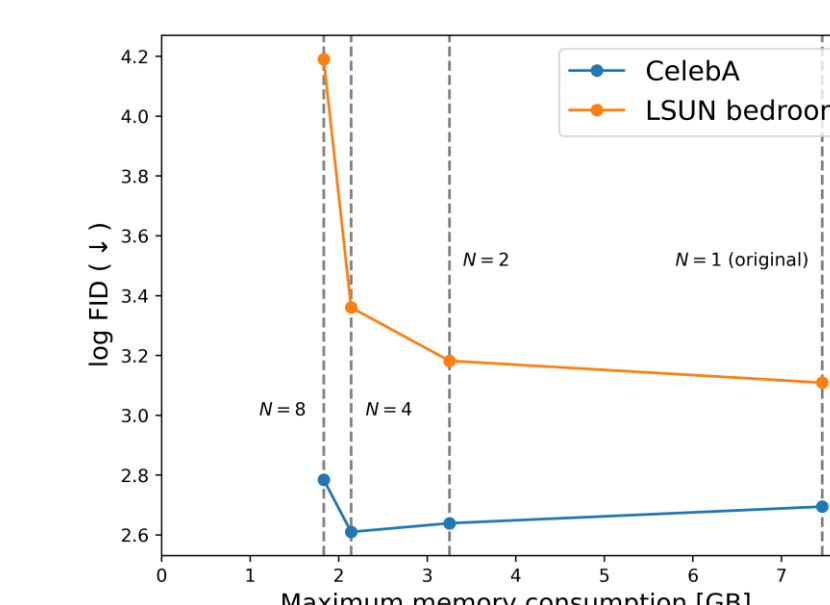
LSUN bedroom^[5]



FID scores

Dataset	Division N	FID ($\downarrow$)
CelebA	1 (baseline)	14.8
	2	14.0
	4	13.6
	8	16.2
LSUN bedroom	1 (baseline)	22.4
	2	24.1
	4	28.8
	8	66.1

Memory Consumption



Summary

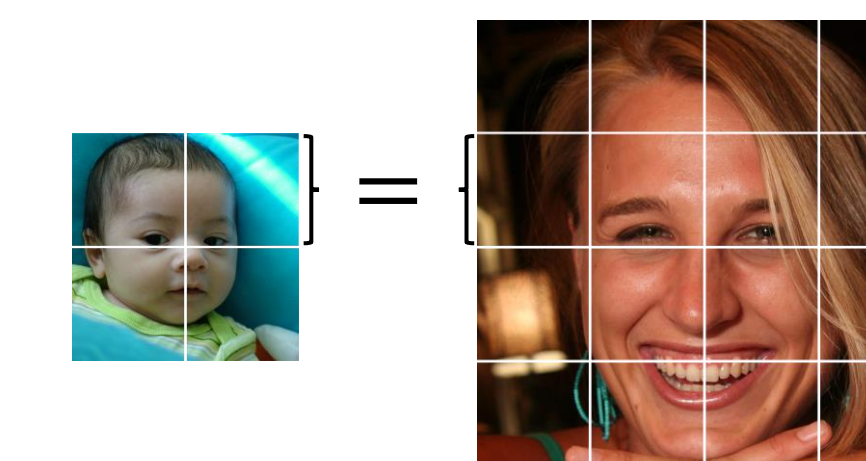
- **Quite natural images when $N = 2, 4$**
- There are discontinuities on patch boundaries when $N = 8$ (mainly on LSUN)
- In $N = 8$, the FID scores of LSUN bedroom are degraded sharply
 - This is because LSUN contains images with **wide variety of compositions**

Conclusions & Future Work

- Patch-by-patch image generation to reduce memory consumption
- Proposed **Positional Conditioning Method** and **Global Content Conditioning**
- Reduced large amount of memory consumption while maintaining image quality comparable to the baseline DDPM.

Future Work

- Training with multi-resolution datasets
- Generate multi-scale images^[6]



References

- [1] Ho et al., “Denoising diffusion probabilistic models,” NeurIPS, 2020.
- [2] Lin et al., “COCO-GAN: Generation by parts via conditional coordinating,” ICCV, 2019.
- [3] Wang et al., “Patch Diffusion: Faster and More Data-Efficient Training of Diffusion Models,” arXiv:2304.12526, 2019.
- [4] Liu et al., “Deep Learning Face Attributes in the Wild,” ICCV, 2015.
- [5] Yu et al., “LSUN: construction of a large-scale image dataset using deep learning with humans in the loop,” arXiv:1506.03365, 2015.
- [6] Chai et al., “Any-resolution Training for High-resolution Image Synthesis,” ECCV, 2022.

