Arbitrary-Scale Image Generation and Upsampling using Latent Diffusion Model and Implicit Neural Decoder

Supplementary Material

In this Supplementary Material, we describe the evaluation metrics in more details. In addition, we present more qualitative results for comparing other models and visualizing the quality of output images.

S.1. Metrics

FID [5]. Fréchet inception distance is a metric used for evaluating the quality of generated images produced by generative models. It measures the similarity between the distribution of real images and the distribution of generated images by computing the Fréchet distance in feature space.

Precision and Recall [8]. Precision and recall are proposed metrics to evaluate fidelity and diversity. Precision refers to the ratio of the generated image to the real image distribution and refers to the precision of how accurately the generated image depicts the real image sample. Recall refers to the ratio of the actual image to the distribution of the generated image sample and refers to the diversity of the generated image.

SelfSSIM [11]. It is a metric for evaluating the scale consistency of the generated images. We downsample two images of different resolutions to a lower resolution and then measure the SSIM [12] between them.

PSNR. Peak Signal-to-Noise Ratio (PSNR) is a widely used metric to quantify the quality of a reconstructed image compared to the original image. A higher PSNR score indicates a lower distortion. It suggests that the processed image is closer to the original in terms of pixel-wise similarity. However, recent research shows that this metric has limitations in indicating actual perceptual quality [1, 9, 14].

LPIPS [14]. Learned Perceptual Image Patch Similarity (LPIPS) is a metric created to measure the similarity between image patches from a perceptual standpoint. It has been demonstrated to accurately reflect human perception. A low LPIPS score indicates that the patches are perceptually similar.

S.2. Experiment Details

In this section, we describe the target scale at which our model and the comparison models were trained on each task of the experiment.

S.2.1. Image-Generation

For the results in the comparative experiments, we referred to the results of Ntavelis *et al*. [11]. Note ScaleParty, MSPIE, MS-PE were trained at larger resolutions, e.g. over 256×256 and 128×128 for FFHQ and LSUN respectively, while our method was trained at less than those resolutions.

FFHQ [7]. For the human face generation task, each model generated images at five different scales, 256, 320, 384, 448 and 512. The training policy for each model is as follows.

- MS-PE is trained for every scale in comparison (*i.e*. 256, 320, 384, 448 and 512), since it is a multi-scale generation model.
- CIPS is trained on single scale, 256.
- ScaleParty is trained with two different resolutions, 256 and 384, for its scale consistency approach.
- Our model is trained to generate an image of arbitrary resolution between (32, 256] from a latent vector of 32.

LSUN [13]. For generic scene (bedroom and church) generation tasks, each model generated images at three different scales, 128 160 and 192. The training policy for each model is as follows.

- MSPIE and ScaleParty are trained for 128 and 192.
- Our model is trained to generate an image of arbitrary resolution between (64, 128] from a latent vector of 64.

S.2.2. Super-Resolution

In the super-resolution operation, 16×16 low-resolution images are upsampled to arbitrary scales. All models were trained within a scale range of 8× for human faces and 16× for generic scenes.

S.3. More Results

S.3.1. Quantitative Results

Tab. S1 show additional quantitative results of image generation for LSUN Church. In the generation task, the maximum resolution of the images used by our model for training is lower than that of other models, as mentioned in Sec. S.2.1. Nevertheless, as shown in Fig. 4 and Tab. 1 of the main text and Tab. S1, our model shows competitive results. In particular, our model shows great strengths in terms of diversity and scale consistency. And in the super-resolution task, our model achieves significantly better performance not only in terms of fidelity but also in terms of perceptual quality. All methods were trained at the same scales for the super-resolution task.

S.3.2. Qualitative Results

To demonstrate the performance of our model, we provide more provide more generated images and comparison

Dataset:	LSUN Church						
Method	Res	FID↓	Prec↑	Rec↑	SelfSSIM (5k)↑		
MSPIE	128	**6.67**	**71.95**	44.59	1.00	0.32	0.43
	160	10.76	66.21	36.95	0.31	1.00	0.40
	192	**6.02**	66.70	46.13	0.39	0.38	1.00
Scaleparty	128	9.08	70.52	39.93	1.00	0.95	0.93
	160	**7.96**	**70.87**	32.07	0.94	1.00	0.95
	192	7.52	**68.14**	33.33	0.90	0.94	1.00
Ours	128	8.25	65.27	**47.02**	1.00	**0.98**	**0.98**
	160	8.58	64.02	**43.04**	**0.97**	1.00	**0.99**
	192	8.81	62.36	**42.80**	**0.96**	**0.97**	1.00

Table S1. Quantitative comparison of image generation on LSUN Church datasets.

results. In Figs. S1 to S5, we visualize various randomly sampled results for the FFHQ, LSUN-Bedroom and LSUN-Church datasets, respectively. Our model shows remarkable performance in synthesizing high-quality details with a variety of styles and scale-consistency.

Figs. S6 to S8 show the qualitative comparison of *SR* for CelebA-HQ [6], LSUN-Bedroom and LSUN-Tower, respectively. LIIF has over-smoothing issues in contrast to high PSNR scores. Both IDM and our model are good at capturing high-resolution details, and furthermore, our model has achieved relatively few distortions. In addition, Fig. S9 shows various *SR* results for the LSUN datasets. The top image is an *LR* image, and the images below are different *SR* results in the red area. As the scale increases, the number of high-resolution solutions that can be recovered from low-resolution becomes more diverse. However, INR-based models such as LIIF [3] always achieve only the same results. In contrast, our stochastic model can generate a variety of patterns and textures for blankets, clouds, and buildings, etc. while maintaining *LR* information. This allows our model to better handle the 'ill-posed problem'.

References

[1] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *CVPR*, 2018. 1

[2] Kelvin C.K. Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *CVPR*, pages 14245–14254, 2021.

[3] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *CVPR*, pages 8628–8638, 2021. 2

[4] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *CVPR*, pages 10021–10030, 2023.

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017. 1

[6] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *ICLR*, 2018. 2

[7] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1

[8] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *NeurIPS*, 32, 2019. 1

[9] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017. 1

[10] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, 2020.

[11] Evangelos Ntavelis, Mohamad Shahbazi, Iason Kastanis, Radu Timofte, Martin Danelljan, and Luc Van Gool. Arbitrary-scale image synthesis. In *CVPR*, pages 11533–11542, 2022. 1

[12] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1

[13] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 1

[14] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1

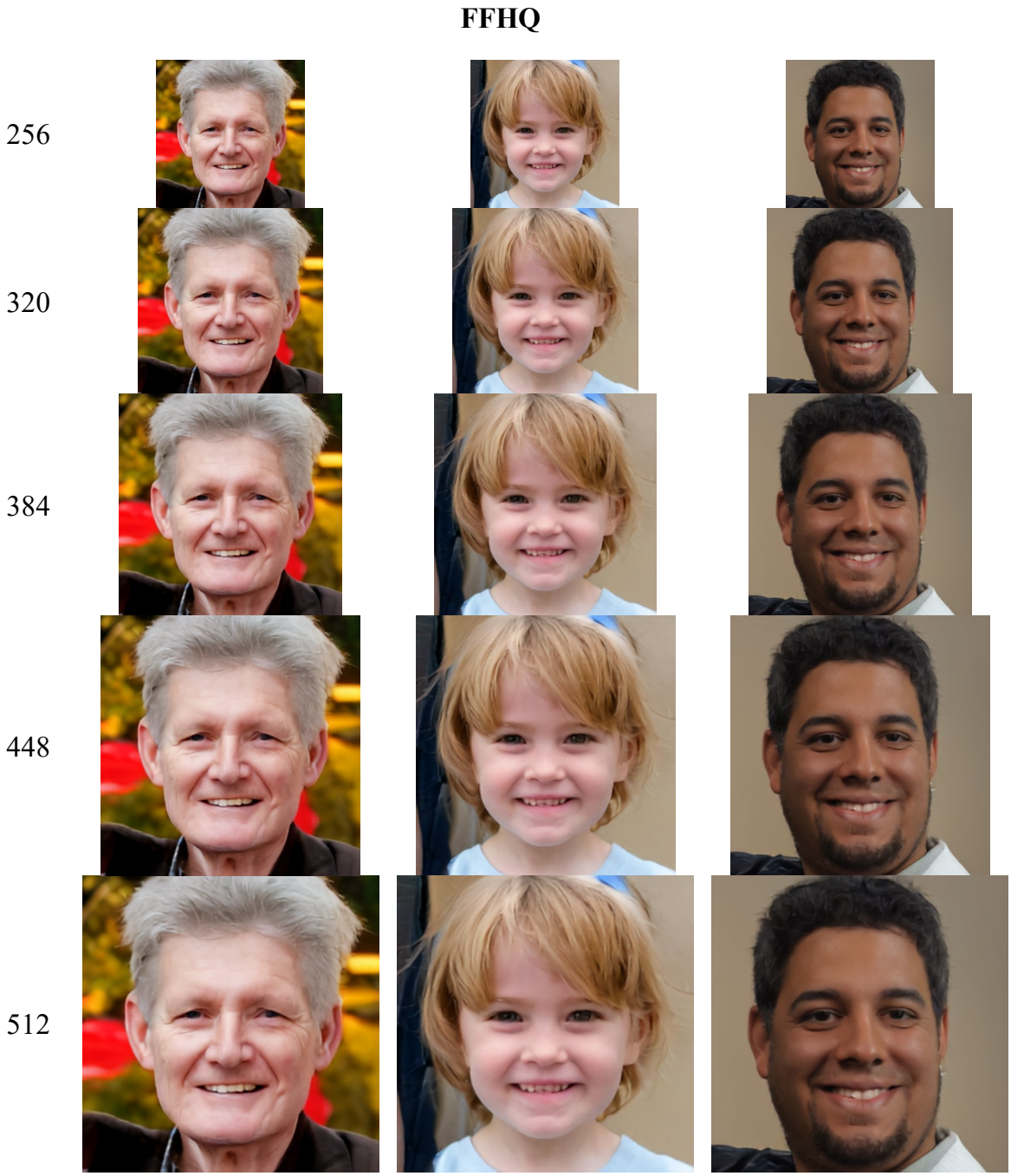

Figure S1. Scale consistency results of image generation on the FFHQ datasets.

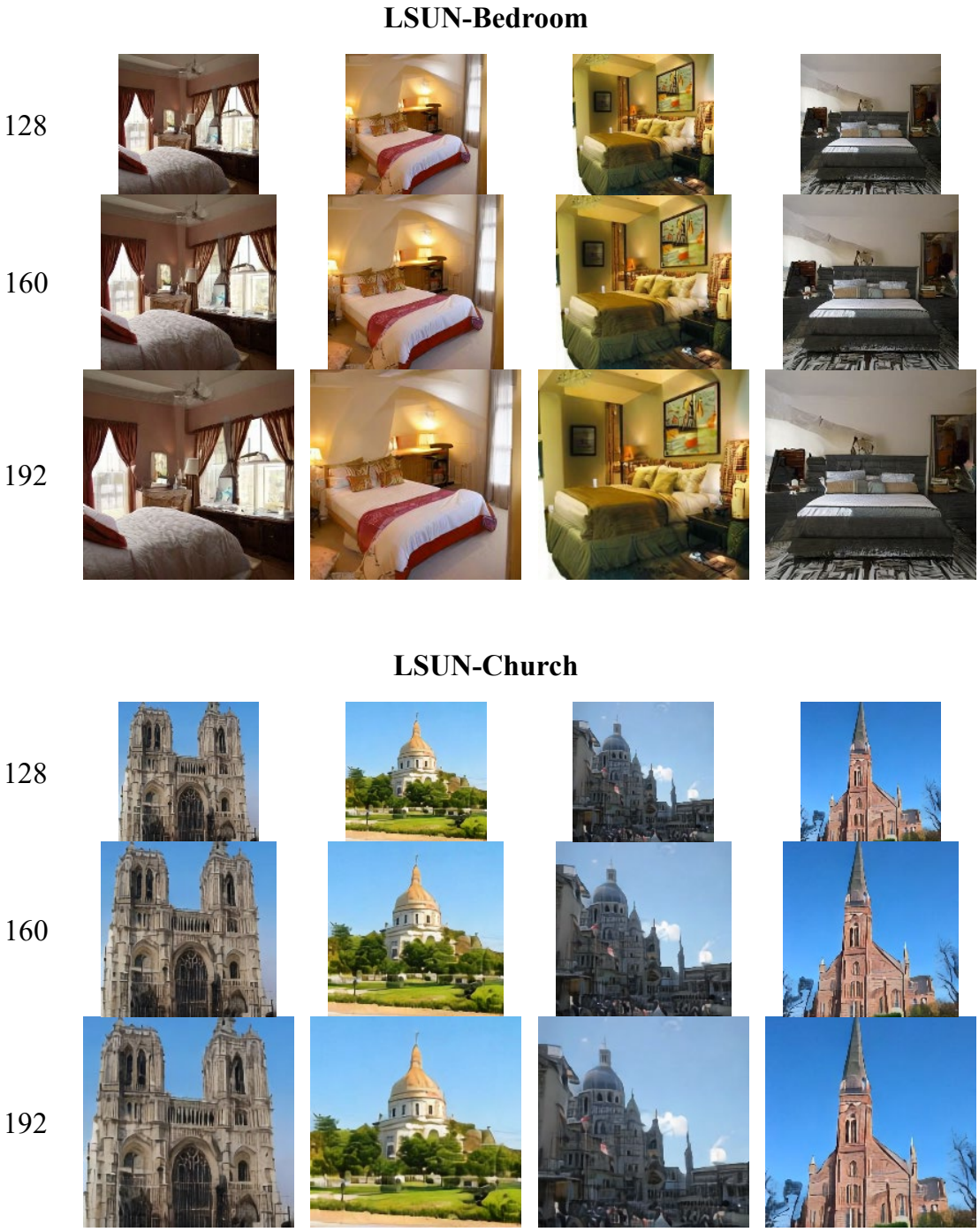

Figure S2. Scale consistency results of image generation on the LSUN Bedroom, Church datasets.

FFHQ

Figure S3. Visual results of image generation on the FFHQ datasets.

LSUN-Bedroom

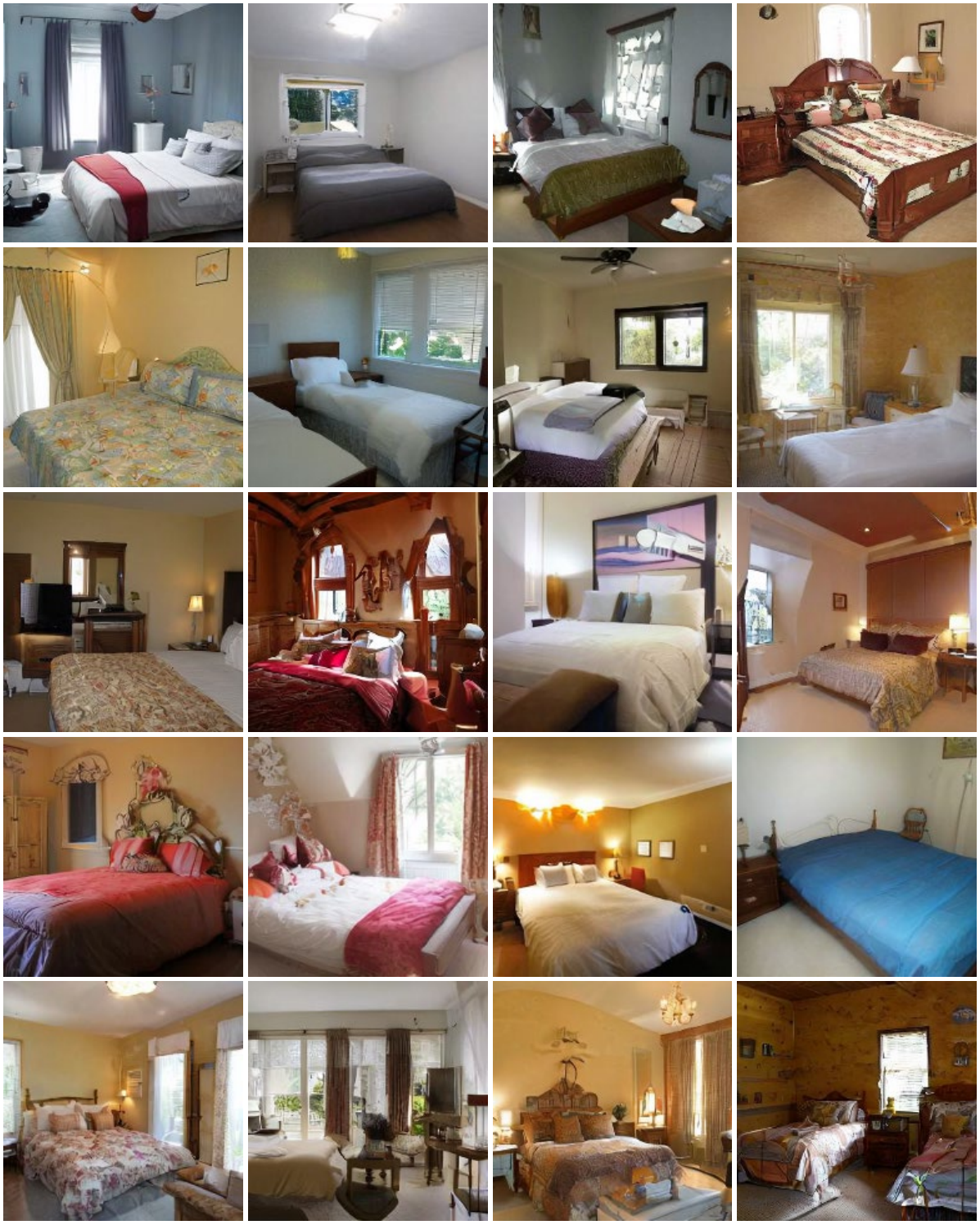

Figure S4. Visual results of image generation on the LSUN Bedroom datasets.

LSUN-Church

Figure S5. Visual results of image generation on the LSUN Church datasets.

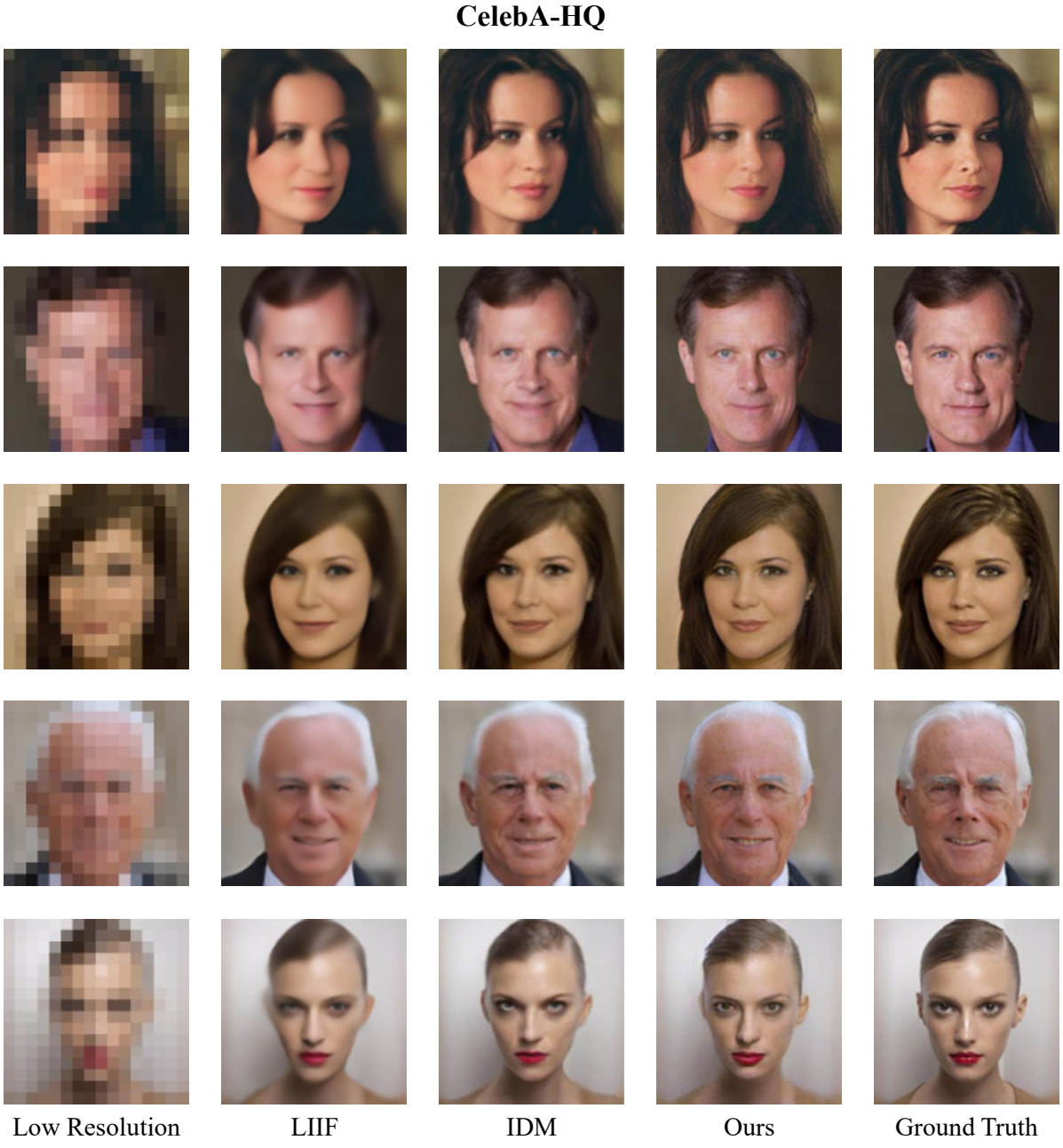

Figure S6. Comparison of $16 \times 16 \rightarrow 128 \times 128$ super-resolution on the CelebA-HQ datasets.

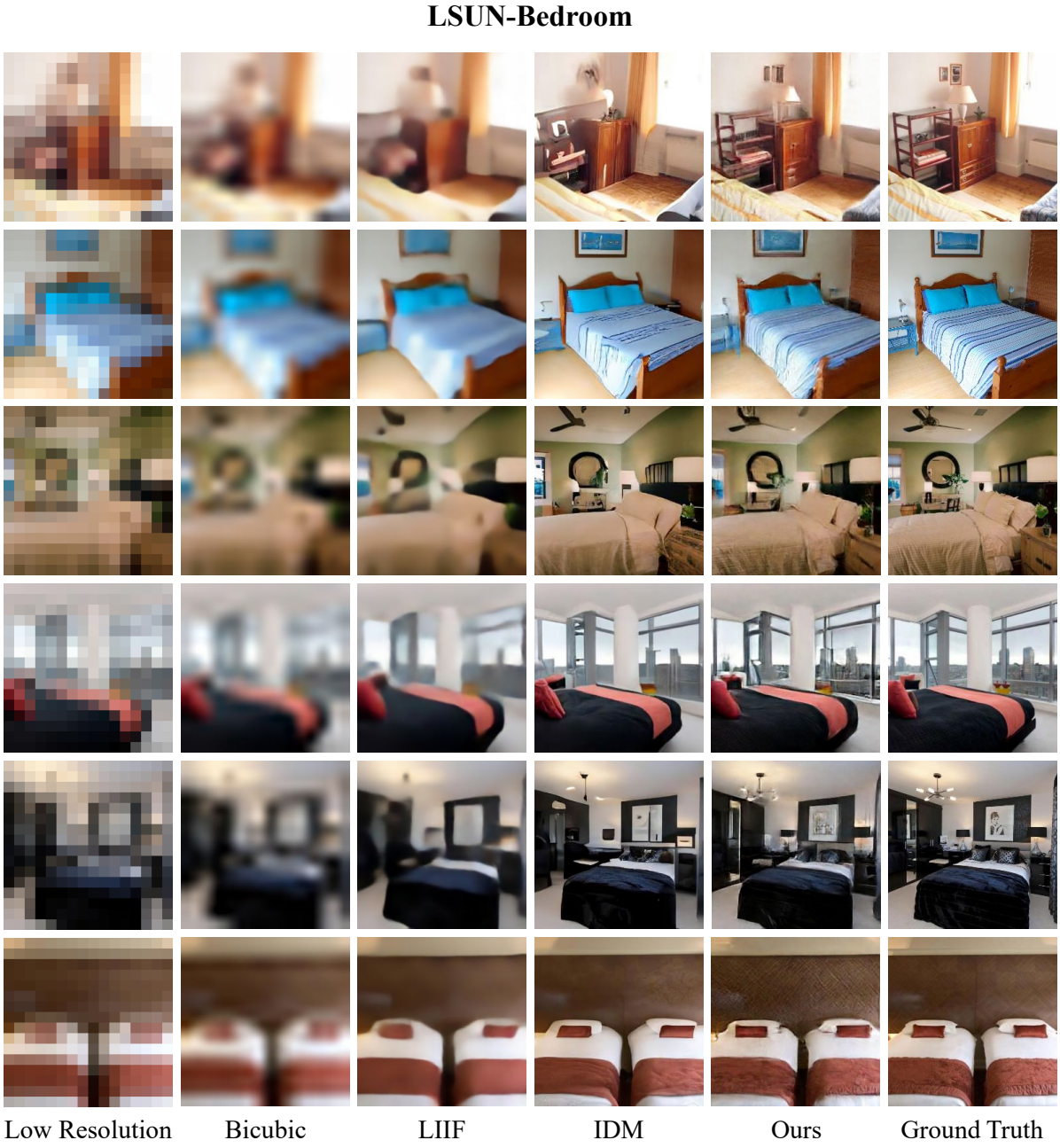

Figure S7. Comparison of $16 \times 16 \rightarrow 256 \times 256$ super-resolution on the LSUN Bedroom datasets.

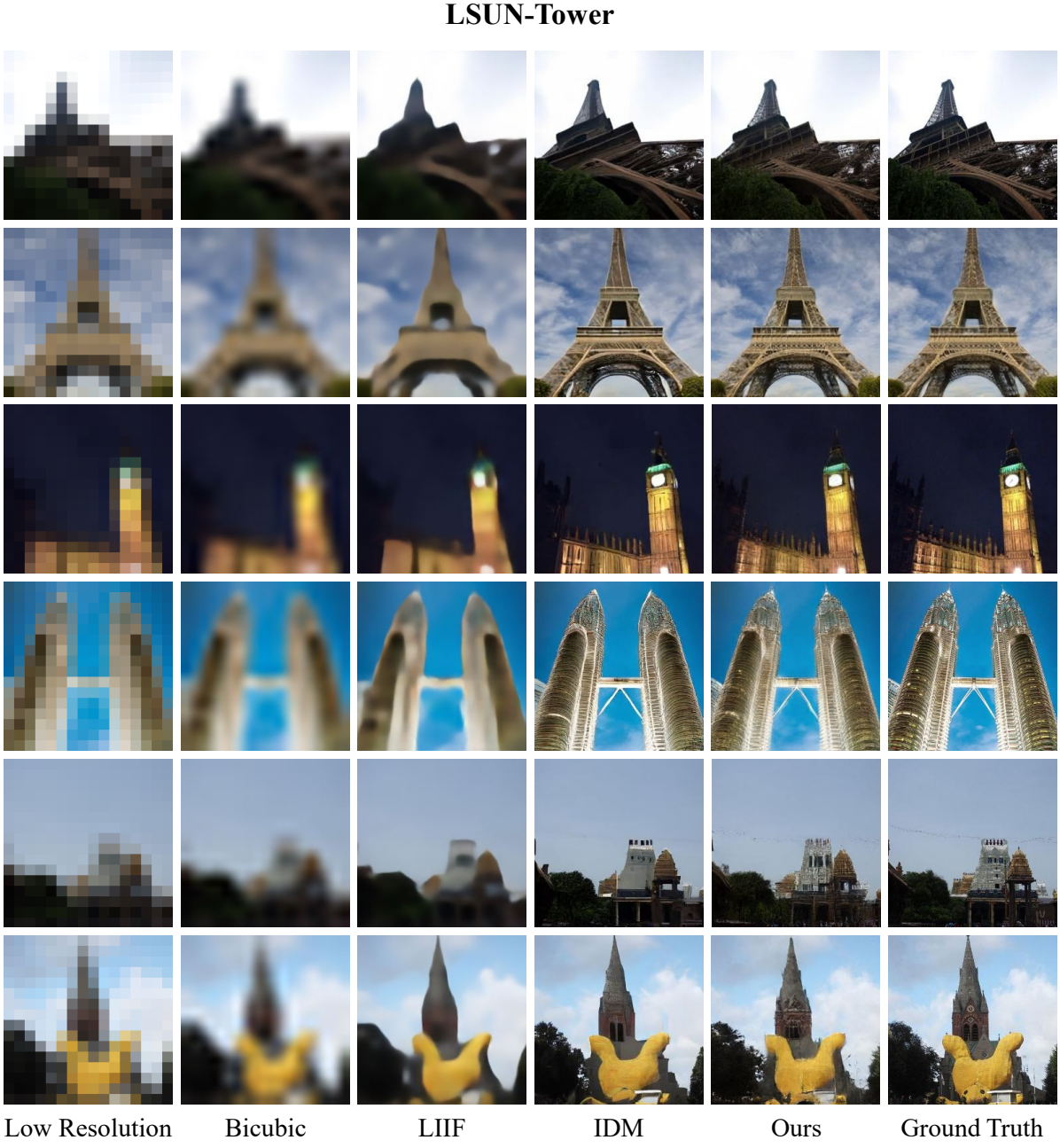

Figure S8. Comparison of $16 \times 16 \rightarrow 256 \times 256$ super-resolution on the LSUN Tower datasets.

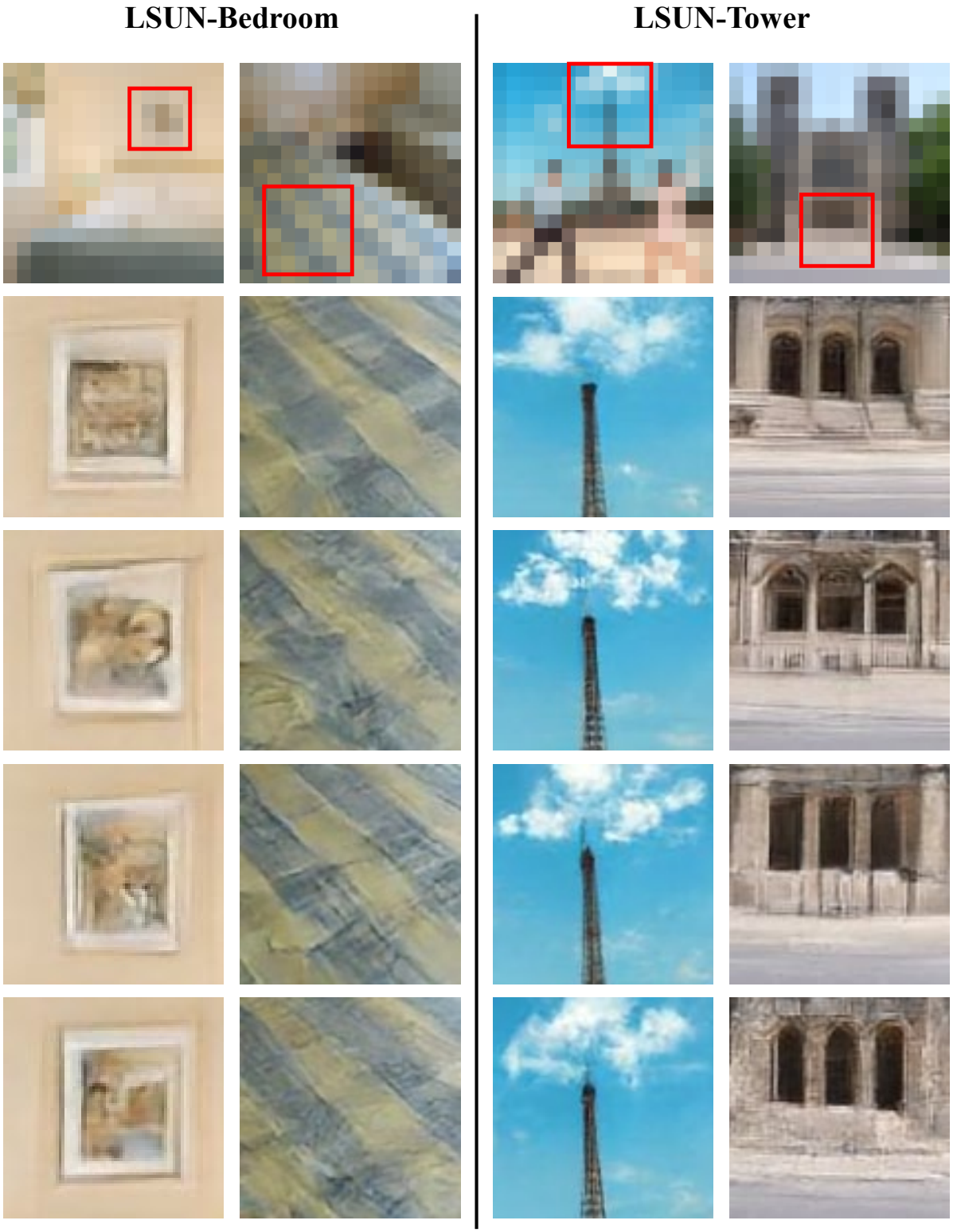

Figure S9. Visualization results of diversity in super-resolution tasks. The top image is an *LR* image, and the images below are different *SR* results in the red area.