Supplementary Material

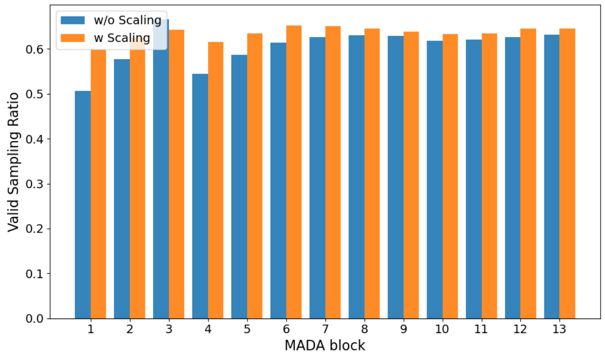

Figure A: Comparison of valid sampling ratios across the Mask-Aware Deformable Attention (MADA) blocks when trained with and without Adaptive Offset Range Scaling.

A. Implementation Details

Implementation Details. Mask-Aware Deformable Inpainting Network (MADIN) consists of a total of 7 stages of transformer blocks, with the number of MADA modules per stage set to $[1, 2, 4, 6, 4, 2, 1]$ in order. The spatial resolutions of the feature maps across these stages are given by $[H\times W, \frac{H}{2}\times\frac{W}{2}, \frac{H}{4}\times\frac{W}{4}, \frac{H}{8}\times\frac{W}{8}, \frac{H}{4}\times\frac{W}{4}, \frac{H}{2}\times\frac{W}{2}, H\times W]$, where downsampling and upsampling between stages are performed using convolutional layers. In all MADA blocks, the number of sampling points is fixed at 64×64, which is maintained by adjusting the stride s following the setting of DAT. All training is performed at an input resolution of 256×256, and inference is also conducted at 256×256 except for the real-world application experiments. On CelebA-HQ, the model is trained for 878k iterations, while on Places2 it is trained for 2705k iterations. All experiments are conducted on a single NVIDIA RTX 4090 GPU with a batch size of 4. The AdamW optimizer is used, with hyperparameters set to $\beta_1 = 0.5$, $\beta_2 = 0.9$.

B. Adaptive Offset Range Scaling

We compared models trained with and without Adaptive Offset Range Scaling by measuring the ratio of sampling locations that fall into valid regions, as summarized in Figure A. MADIN has seven stages with multiple Mask-Aware Deformable Attention (MADA) blocks, and we averaged results across blocks within each stage, using the 13 encoder blocks for visualization. The evaluation was performed at an input resolution of 256×256 on 100 test images from the Places2 dataset, each combined with 100 masks, where the average masked area was 0.4578.

The model trained with Adaptive Offset Range Scaling showed consistently higher valid sampling ratios across almost all blocks, with an average improvement of about 3.98 percentage points. Since Adaptive Offset Range Scaling introduces neither additional parameters nor noticeable computational cost, it can be considered an efficient auxiliary component for guiding offsets toward valid regions.

C. Additional Qualitative Results

We provide additional qualitative comparisons on CelebA-HQ and Places2 at a resolution of 256×256. The examples span diverse mask patterns and object categories, including cases with large missing regions and complex structural details. For each masked input, we juxtapose the outputs of recent transformer-based approaches (MAT, CMT, and M×T, Latent Codes) and our approach, following the experimental setup described in the main paper.

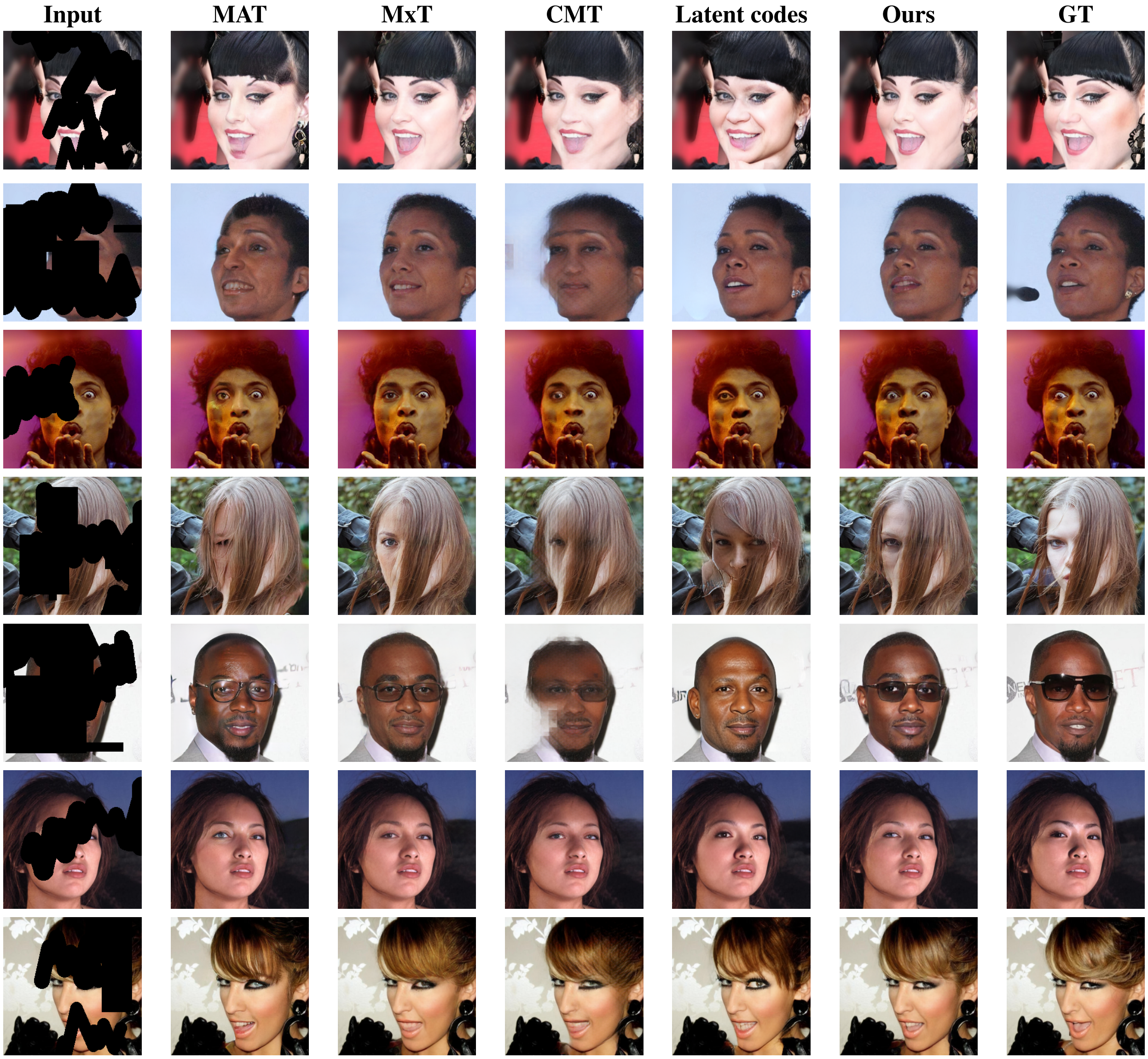

Figure B: Additional qualitative comparison of inpainted results on CelebA-HQ dataset.

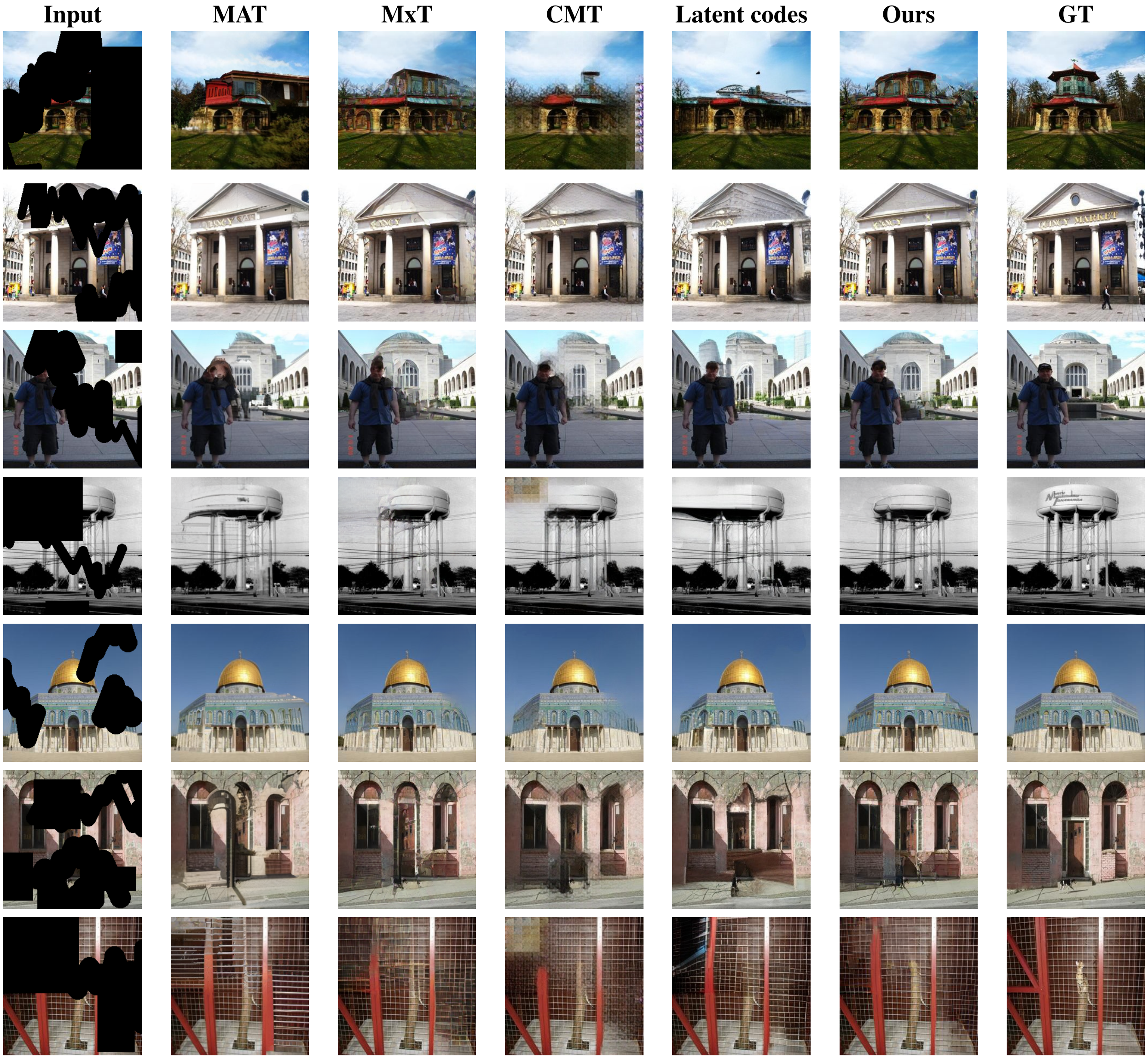

Figure C: Additional Qualitative comparison of inpainted results on Places2 dataset.