# Appendix

## A More Experiments

| Resolution | PES | PIPE | PSF | KID↓ | $KID_p$↓ | IS↑ | $IS_p$↑ | CLIP↑ |
|---|---|---|---|---|---|---|---|---|
| 3072×3072 | × | × | × | 0.0836 | 0.1958 | 7.80 | 4.06 | 25.08 |
| | × | ✓ | ✓ | 0.0906 | 0.0842 | 11.05 | 6.97 | 30.72 |
| | ✓ | × | ✓ | 0.2218 | 0.4426 | 8.54 | 3.03 | 24.17 |
| | ✓ | ✓ | × | 0.0227 | 0.0336 | 12.04 | 9.84 | 32.51 |
| | ✓ | ✓ | ✓ | **0.0189** | **0.0199** | **12.91** | **10.87** | **32.85** |
| 4096×4096 | × | × | × | 0.2703 | 0.3164 | 6.9 | 3.53 | 21.28 |
| | × | ✓ | ✓ | 0.1334 | 0.1239 | 10.16 | 6.56 | 26.63 |
| | ✓ | × | ✓ | 0.2208 | 0.3976 | 8.56 | 3.77 | 24.69 |
| | ✓ | ✓ | × | 0.0488 | 0.0760 | 10.35 | 8.10 | 31.03 |
| | ✓ | ✓ | ✓ | **0.0217** | **0.0252** | **11.46** | **9.97** | **32.71** |

Table 1: **Quantitative ablation study** of PE Scaling (PES) / Patch-wise Independent PE (PIPE) / Patch-wise Spectral Fusion (PSF) components at different resolutions. The best results are marked in **bold**.

| Method | KID ↓ | IS ↑ | CLIP ↑ |
|---|---|---|---|
| MOP × GWS × | 0.0189 | 12.91 | 32.81 |
| MOP ✓ GWS × | 0.0189 | 12.90 | 32.85 |
| ResDiT | 0.0189 | 12.91 | 32.85 |

Table 2: **Quantitative ablation study** of Minimum-Overlap Partitioning (MOP) / Gaussian Weighting Splicing (GWS) components at 3072 × 3072 resolutions.

| Resolution | Metric | Vanilla | Imax | HiFlow | ResDiT |
|---|---|---|---|---|---|
| 3072×3072 | Time (s) | 542 | 568 | **203** | <u>448</u> |
| 4096×4096 | Time (s) | 1451 | 1563 | **465** | <u>1032</u> |

Table 3: **Inference efficiency comparison** of different methods at 3072 × 3072 and 4096 × 4096 resolutions. The fastest method is highlighted in **bold**, and ResDiT is underlined.

We conducted quantitative experiments to further validate the contributions of the three core components in our pipeline, as shown in table 1, using the same implementation details described in the main text. As discussed earlier,

omitting Patch-wise Spectral Fusion mainly causes partial blurring and repeated artifacts in the generated images, primarily affecting visual quality and resulting in a slight drop in quantitative performance compared with ResDiT. In contrast, removing the other two components severely degrades image quality and leads to significantly lower scores across metrics.

Notably, for some metrics, the performance without Patch-wise Independent PE is even worse than direct generation. We believe this is because, although direct generation produces structurally disordered results, it still preserves some local semantic information. By contrast, removing Patch-wise Independent PE yields roughly correct global structures but introduces substantial noise and artifacts throughout the image, which can heavily affect certain metrics.

We also conducted quantitative experiments to evaluate the Minimum-Overlap Partitioning and Gaussian Weighting Splicing components at $3072 \times 3072$ resolution. As expected, these two methods mainly help alleviate boundary stitching artifacts, improving visual quality, while having relatively little impact on the quantitative metrics.

We also evaluate the inference efficiency of different methods at high resolutions, as shown in table 3. ResDiT achieves a favorable trade-off between speed and quality, being significantly faster than Vanilla and Imax at both 3K and 4K resolutions, while slightly slower than the fastest method, HiFlow. We acknowledge that this paradigm can be slower in latency, since HiFlow only requires a few timesteps at high resolution. In terms of peak VRAM, all methods are largely dominated by the maximum-resolution stage and are thus similar.

# B More Qualitative Results

We present additional ResDiT generation results below. Samples at resolution of $3072 \times 3072$ are shown in fig. 1, and results at $4096 \times 4096$ resolution are displayed in fig. 2.

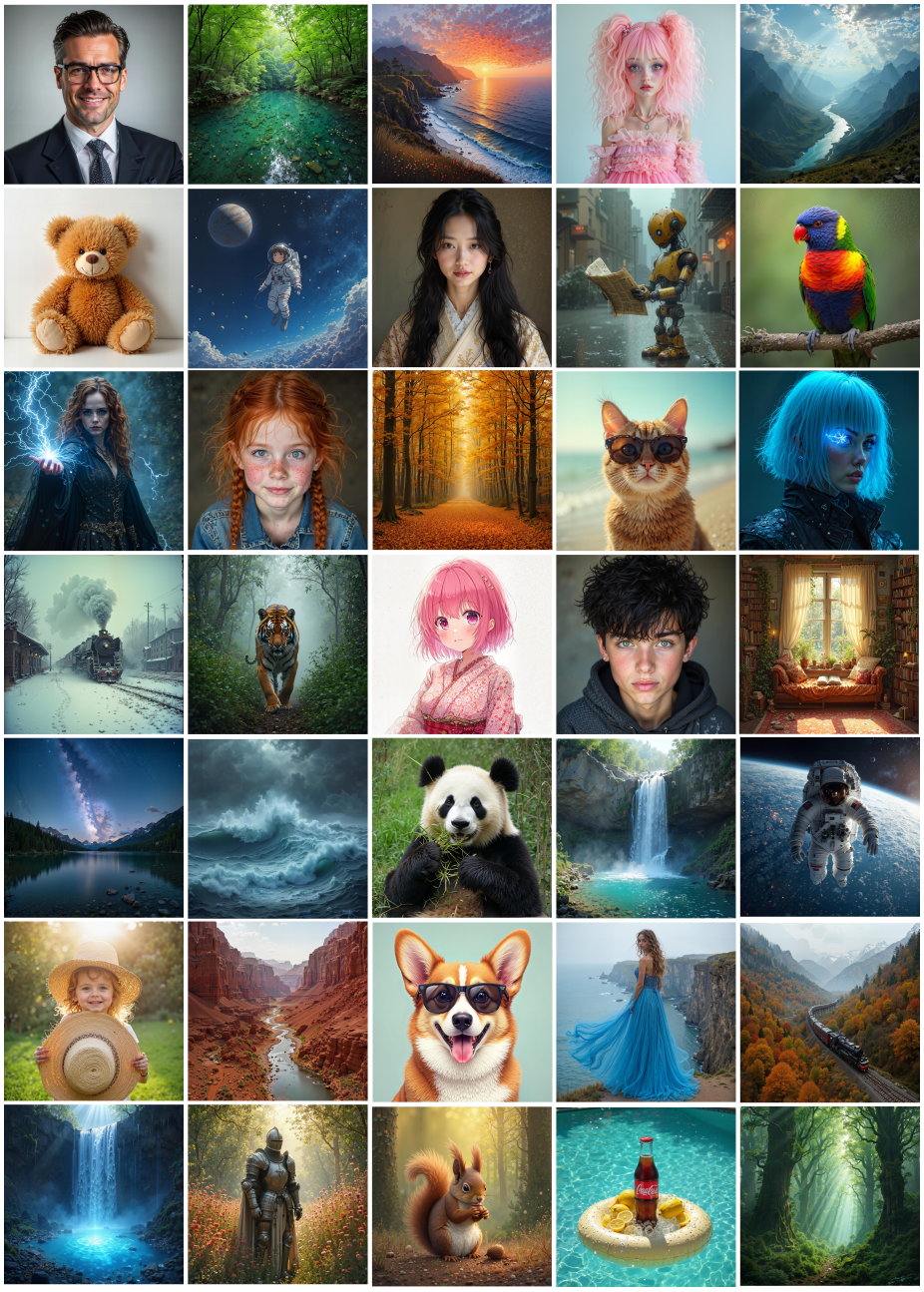

Figure 1: The 3072 × 3072 resolution image generated by ResDiT. **Best View ZOOM-IN.**

Figure 2: The 4096 × 4096 resolution image generated by ResDiT. **Best View ZOOM-IN.**