Otil: Accelerating Diffusion Model Inference via Communication-Efficient Multi-GPU Parallelism

Supplementary Materials

In this work, we provide the following supplementary materials:

• Sec 1 provides comprehensive studies on the quality analysis in Otil.

• Sec 2 presents additional experimental results evaluating the compatibility of Otil with various few-step samplers.

• Sec 3 reports comparative experiments between Otil and other parallel inference frameworks.

• Sec 4 provides comprehensive studies on the sub-block selection mechanism used in Otil.

1. Quality analysis.

Our comprehensive analysis of metric distributions reveals that the vast majority of generated samples maintain high quality: 66.4% of LPIPS values and 73.3% of FID values are concentrated within the optimal range, with only $10-15\%$ of samples showing any deviation. Only 10.2% of PSNR values are lower 21. Our evaluation across the entire dataset shows variance: $\sigma^2_{\text{PSNR}} = 9.94$, $\sigma^2_{\text{LPIPS}} = 0.00067$, and $\sigma^2_{\text{FID}} = 0.98$. This indicates that quality degradation is not a systematic issue but rather a rare edge case affecting a small minority of samples. More importantly, by examining the visual samples with the lowest PSNR, LPIPS, and FID scores, it is evident that these quality differences have a negligible impact on human observers. In Figure 1, we present visual comparisons of the worst-performing samples. The image details are highlighted with red and blue boxes to better observe the differences.

2. The Compatibility Of Otil With Various Few-step Samplers.

To evaluate the compatibility of Otil with various few-step samplers, we conduct 15-step sampling experiments using DPM-Solver, UniPC, and Euler on SD 1.5 (512×512 resolution) and SDXL 1.0 (1024×1024 resolution). We compare the performance of Otil integrated with each sampler against the corresponding original models. As shown in Table 2, on a 2-GPU setup, our method achieves a $3.67\times$ speedup on SD 1.5, with only minimal degradation across all evaluation metrics and under a 4-GPU PCIe interconnect, integrating Otil with DPM-Solver yields a $6.60\times$ speedup on SDXL 1.0, while incurring only a negligible degradation in LPIPS. The final image quality remains virtually identical to the baseline, and representative samples are provided in Figure 2.

Method	Generalization			
	Few-step Samplers	Framework		GPU Numbers Agnostic
		U-Net	DiT	
DistriFusion	✓	✓	×	✓
PipeFusion	×	×	✓	✓
CompactFusion	×	×	✓	×
ParaStep	×	×	✓	✓
AsyncDiff	×	✓	✓	✓
Otil (Ours)	✓	✓	✓	✓

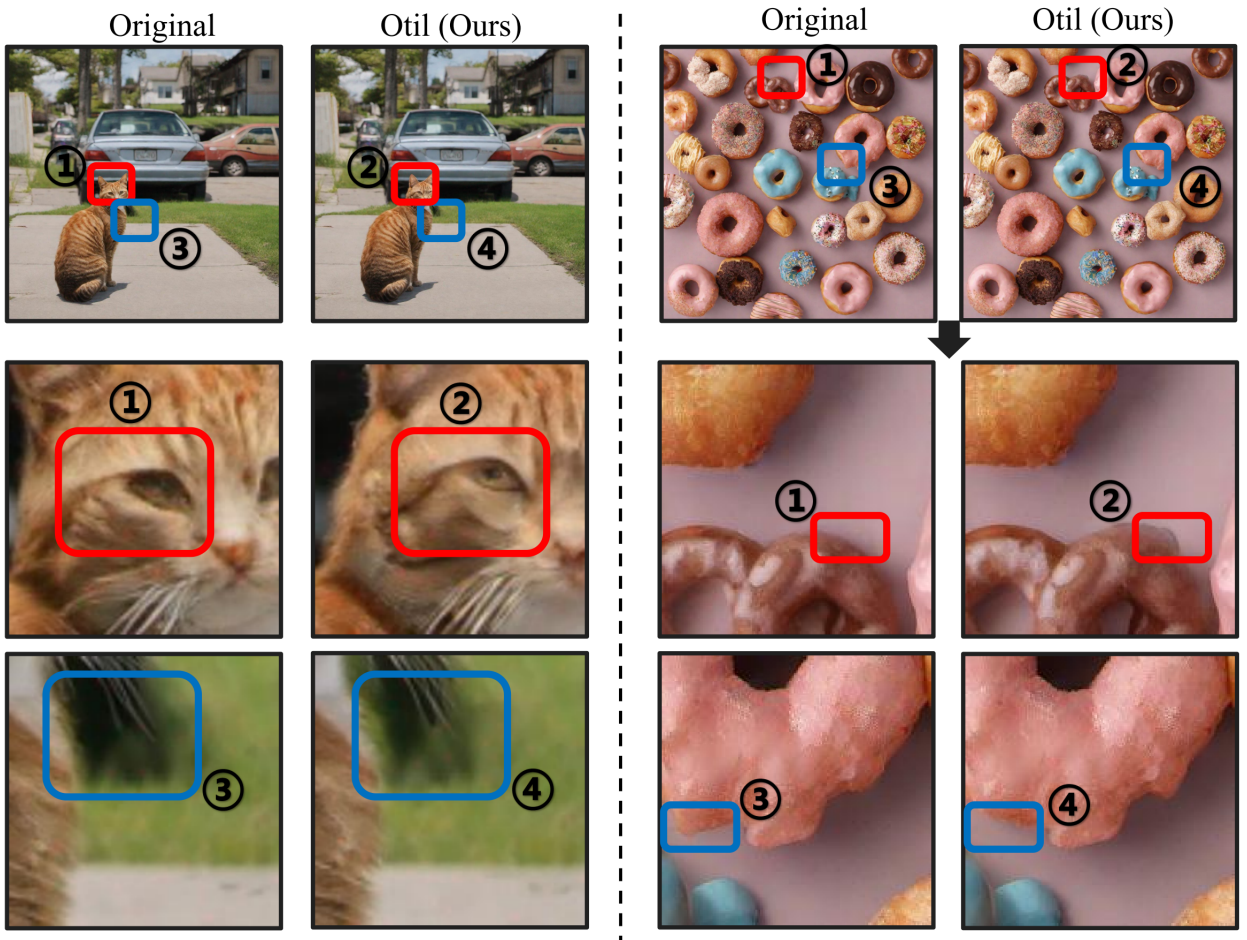
Table 1. Generalization of acceleration methods to different deployment conditions.

3. Compare With Other Parallel Frameworks.

As summarized in Table 1, our original evaluation focused on U-Net backbones (SD1.5, SDXL1.0), where the aforementioned DiT-only methods are not applicable. We instead compared against the strongest U-Net-compatible baselines (e.g., DistriFusion, AsyncDiff).

We conduct a comprehensive comparison between Otil and existing parallel diffusion frameworks. Since AsyncDiff is incompatible with SD 1.5 and does not support few-step samplers, we compare Otil with DistriFusion on both SD 1.5 (512×512 resolution) and SDXL 1.0 (1024×1024 resolution) under two configurations: with and without integrating DPM-Solver. As summarized in Table 3, our method consistently achieves higher speedups than DistriFusion while preserving image quality across all settings. The corresponding performance visualizations are shown in Figure 3.

Furthermore, we generate 1024×1024 images on SDXL 1.0 using 50-step DDIM sampling and evaluate three parallel inference frameworks—DistriFusion, AsyncDiff and Otil (Ours)—on 2 and 4 GPUs connected via PCIe. Representative qualitative results are presented in Figure 4. As shown, our method effectively reduces communication latency while still preserving the final image quality.



(a) Prompt: A cat in between two cars in a parking lot.

(b) Prompt: This is a case of various colored doughnuts.

Figure 1. In both sets of generated images, the left image is produced by the original model, while the right image is generated using the Otil method. A careful visual comparison reveals differences in the fine details of the images, particularly around edges and textures. These subtle variations, while relatively minor to the human observer, contribute to a slight decrease in our image quality scores (e.g., PSNR). Despite these differences, the overall visual quality remains largely preserved, indicating that the impact on perceptual quality is minimal.

4. Sub-block Selection Comprehensive Experiment.

To comprehensively evaluate the impact of the difference-measure computation on the identification of the most informative sub-blocks, we conduct extensive experiments across multiple configurations and settings. Specifically, we assess the performance under the following conditions:

- Sub-block size of 4×4, 8×8 and 16×16.
- Difference-measure algorithms including cosine similarity, SSIM (not applicable to 4×4 sub-block size), MI, dHash, and random selection.
- The selection ratios are set to $\frac{1}{2}$, $\frac{1}{4}$ and $\frac{1}{8}$ of the total number of blocks.

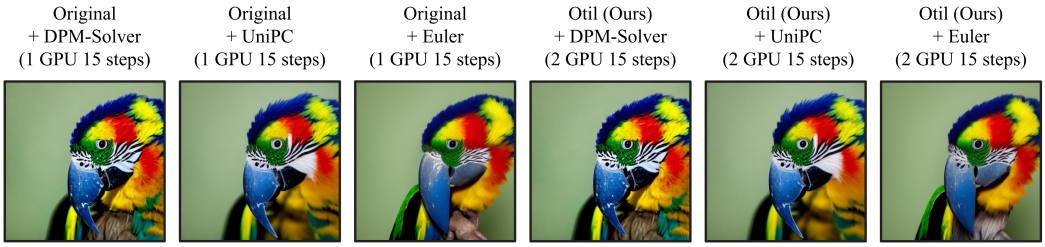
As shown in Figure 5, we evaluate the identification performance of five sub-block selection algorithms under various sub-block sizes and different selection ratios ($\frac{k}{K} = \frac{1}{2}, \frac{1}{4}, \frac{1}{8}$). For each experimental setting, we decode the intermediate outputs at each denoising step and compute region-wise differences between consecutive steps in both the latent space and the pixel space separately. We then rank the regions in both spaces, ensuring that we account for the variations that occur at each step of the denoising process. The consistency between the rankings from the latent space and pixel space is then quantitatively measured using the area under the curve (AUC) metric. This process enables us to systematically assess the alignment between the rankings from the two different spaces. Across all experimental configurations, cosine similarity consistently demonstrates the best identification accuracy, establishing it as the criterion for selecting the most informative sub-blocks in the latent space.

Number of Devices	Base Model	method	Lantency		PSNR↑	LPIPS↓	FID↓	CLIP Score↑
			Value(s)↓	Speedup↑				
1	SD 1.5	Original + DPM-Solver (15 Steps)	0.567	2.43	21.117	0.191	30.650	31.437
		Original + UniPC (15 Steps)	0.598	2.31	21.119	0.191	30.637	31.441
		Original + Euler (15 Steps)	0.586	2.35	20.876	0.244	31.187	31.097
	SDXL 1.0	Original + DPM-Solver (15 Steps)	1.764	3.35	22.317	0.105	25.511	33.116
		Original + UniPC (15 Steps)	1.794	3.29	22.303	0.109	25.507	33.112
		Original + Euler (15 Steps)	1.784	3.31	20.297	0.215	28.430	29.154
2	SD 1.5	Otil (Ours) + DPM-Solver (15 Steps)	0.375	3.67	20.330	0.269	31.110	31.242
		Otil (Ours) + UniPC (15 Steps)	0.386	3.58	20.227	0.271	31.114	31.242
		Otil (Ours) + Euler (15 Steps)	0.389	3.55	20.032	0.275	32.093	31.037
	SDXL 1.0	Otil (Ours) + DPM-Solver (15 Steps)	1.056	5.59	22.197	0.118	26.332	33.960
		Otil (Ours) + UniPC (15 Steps)	1.062	5.56	22.293	0.109	25.913	32.973
		Otil (Ours) + Euler (15 Steps)	1.052	5.61	20.293	0.219	28.712	29.154
4	SDXL 1.0	Otil (Ours) + DPM-Solver (15 Steps)	0.895	6.60	20.227	0.193	32.122	32.285
		Otil (Ours) + UniPC (15 Steps)	0.901	6.56	20.212	0.193	32.157	32.281
		Otil (Ours) + Euler (15 Steps)	0.912	6.48	19.107	0.276	33.134	31.117

Table 2. The Compatibility Of Otil With Various Few-step Samplers.

Base Model	Number of Devices	Method	Latency		PSNR↑	LPIPS↓	FID↓	CLIP Score↑
			Value(s)↓	Speedup↑				
SD 1.5	2	DistriFusion + DPM-Solver (15 Steps)	0.394	3.5	20.371	0.257	31.110	31.251
		Otil (Ours) + DPM-Solver (15 Steps)	**0.375**	**3.67**	**20.33**	**0.269**	**31.110**	**31.242**
		DistriFusion + DPM-Solver (30 Steps)	0.654	2.11	21.580	0.191	29.520	31.425
		Otil (Ours) + DPM-Solver (30 Steps)	**0.645**	**2.14**	**21.565**	**0.197**	**29.570**	**31.425**
SDXL 1.0	2	DistriFusion + DPM-Solver (15 Steps)	1.106	5.34	22.214	0.115	26.177	33.971
		Otil (Ours) + DPM-Solver (15 Steps)	**1.056**	**5.59**	**22.197**	**0.118**	**26.332**	**33.960**
		DistriFusion + DPM-Solver (30 Steps)	2.122	2.78	24.530	0.098	24.962	33.205
		Otil (Ours) + DPM-Solver (30 Steps)	**2.112**	**2.79**	**24.511**	**0.098**	**24.997**	**33.204**
	4	DistriFusion + DPM-Solver (15 Steps)	0.926	6.38	20.237	0.189	31.887	32.286
		Otil (Ours) + DPM-Solver (15 Steps)	**0.895**	**6.60**	**20.227**	**0.193**	**32.122**	**32.285**
		DistriFusion + DPM-Solver (30 Steps)	1.700	3.47	22.251	0.152	30.032	33.157
		Otil (Ours) + DPM-Solver (30 Steps)	**1.689**	**3.48**	**22.159**	**0.155**	**30.157**	**32.985**

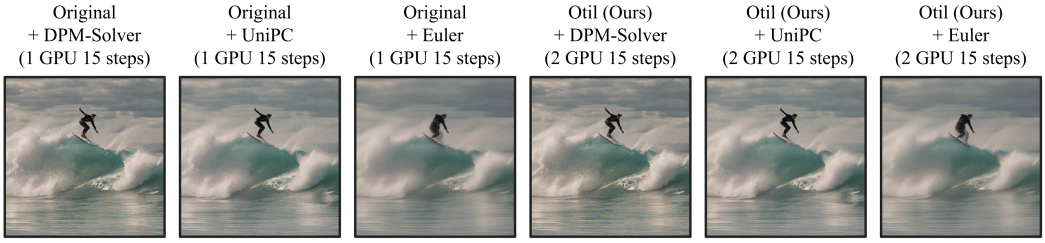
Table 3. Compare with DistriFusion.



Prompt : A multi-colored parrot holding its foot up to its beak.

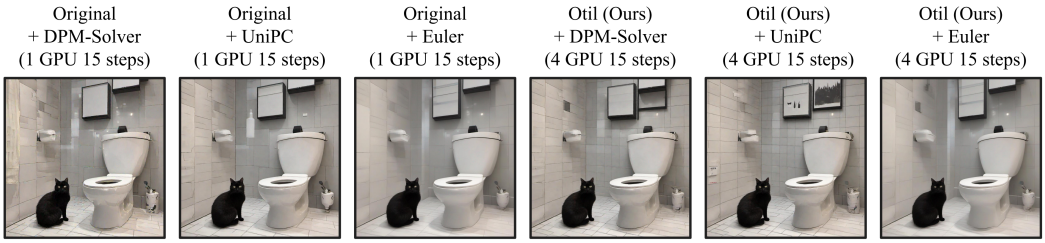
Prompt : Plate of food with gravy on mesh table with knife.

(a) Otil with few-step samplers on SD 1.5



Prompt : A person riding a wave on top of a surfboard.

Prompt : A cup of coffee sits next to a panini sandwich on a counter.

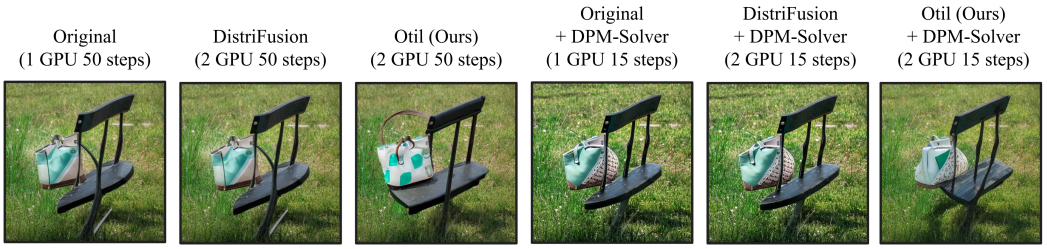


Prompt : A black cat is inside a white toilet.

Prompt : The sign of a restaurant in the outside of the store.

(b) Otil with few-step samplers on SDXL 1.0

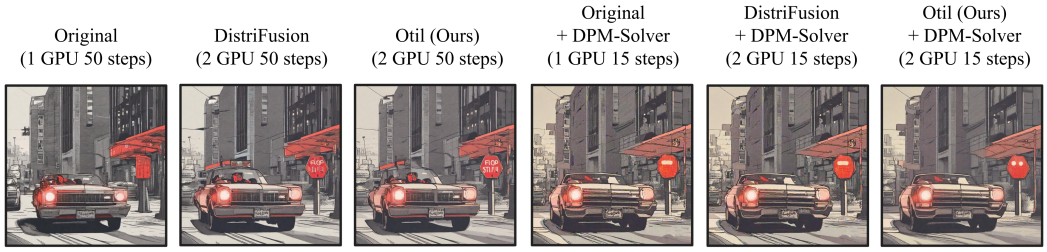
Figure 2. The Compatibility Of Otil With Various Few-step Samplers.



Prompt : A brown purse is sitting on a green bench.

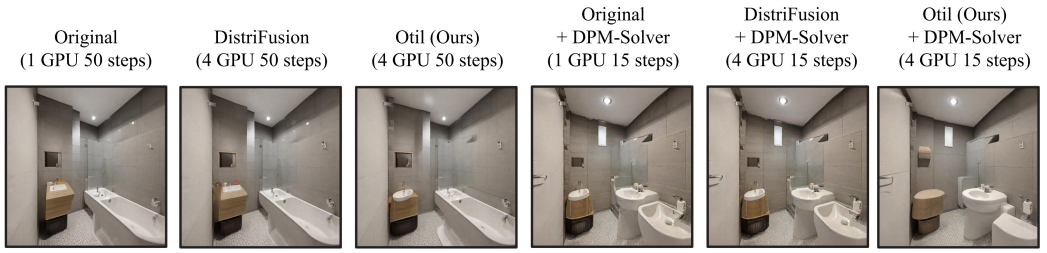
Prompt : The kitchen has a white door with a window.

(a) Compare with DistriFusion on SD 1.5



Prompt : A car is stopped at a red light.

Prompt : A dining table with a bowl of garlic cloves.



Prompt : Small bathroom with a toilet and a sink.

Prompt : A bicycle parked next to a flooded river.

(b) Compare with DistriFusion on SDXL 1.0

Figure 3. Compare With DistriFusion.

Figure 4. Visualisation of Results.

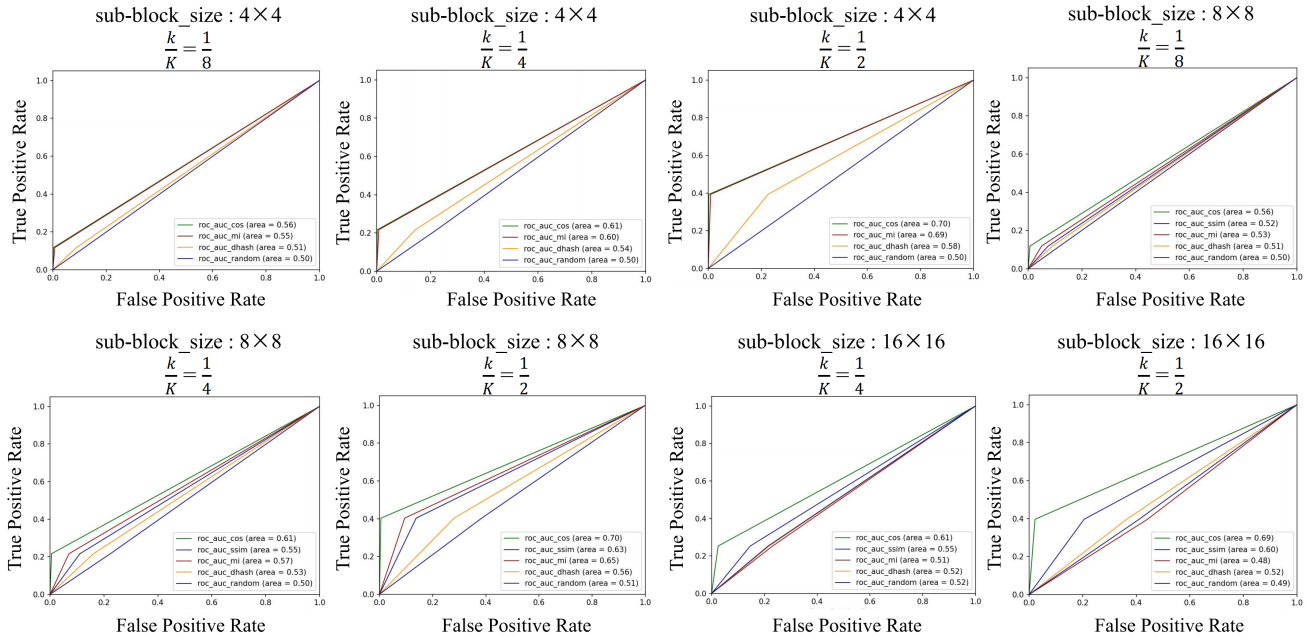


Figure 5. Sub-block Selection Comprehensive Experiment.