DRiffusion: Draft-and-Refine Process Parallelizes Diffusion Models with Ease

Supplementary Material

7. Derivation for Skip Transition

For DDPM:

$$
\begin{aligned}
& q\left(\boldsymbol{x}_{t-k} | \boldsymbol{x}_t, \boldsymbol{x}_0\right) \\
= & \frac{q\left(\boldsymbol{x}_t | \boldsymbol{x}_{t-k}\right) q\left(\boldsymbol{x}_{t-k} | \boldsymbol{x}_0\right)}{q\left(\boldsymbol{x}_t | \boldsymbol{x}_0\right)} \\
= & \frac{\mathcal{N}\left(\boldsymbol{x}_t ; \sqrt{\prod_{i=t-k+1}^t \alpha_i} \boldsymbol{x}_{t-1},\left(1-\prod_{i=t-k+1}^t \alpha_i\right) \mathbf{I}\right) \mathcal{N}\left(\boldsymbol{x}_{t-k} ; \sqrt{\alpha_{t-k}} \boldsymbol{x}_0,\left(1-\alpha_{t-k}\right) \mathbf{I}\right)}{\mathcal{N}\left(\boldsymbol{x}_t ; \sqrt{\alpha_t} \boldsymbol{x}_0,\left(1-\alpha_t\right) \mathbf{I}\right)} \\
\propto & \exp\left\{-\frac{1}{2}\left[\frac{\left(\boldsymbol{x}_t-\sqrt{\frac{\alpha_t}{\alpha_{t-k}}} \boldsymbol{x}_{t-1}\right)^2}{1-\frac{\alpha_t}{\alpha_{t-k}}}+\frac{\left(\boldsymbol{x}_{t-k}-\sqrt{\alpha_{t-k}} \boldsymbol{x}_0\right)^2}{1-\alpha_{t-k}}-\frac{\left(\boldsymbol{x}_t-\sqrt{\alpha_t} \boldsymbol{x}_0\right)^2}{1-\alpha_t}\right]\right\} \\
\propto & \exp\left\{-\frac{1}{2}\left[-\frac{2 \sqrt{\frac{\alpha_t}{\alpha_{t-k}}} \boldsymbol{x}_t \boldsymbol{x}_{t-k}}{1-\frac{\alpha_t}{\alpha_{t-k}}}+\frac{\frac{\alpha_t}{\alpha_{t-k}} \boldsymbol{x}_{t-k}^2}{1-\prod_{i=t-k+1}^t \alpha_i}+\frac{\boldsymbol{x}_{t-k}^2}{1-\alpha_{t-k}}-\frac{2 \sqrt{\alpha_{t-k}} \boldsymbol{x}_{t-k} \boldsymbol{x}_0}{1-\alpha_{t-k}}+C\left(\boldsymbol{x}_t, \boldsymbol{x}_0\right)\right]\right\} \\
\propto & \exp\left\{-\frac{1}{2}\left[\frac{1-\alpha_t}{\left(1-\frac{\alpha_t}{\alpha_{t-k}}\right)\left(1-\alpha_{t-k}\right)} \boldsymbol{x}_{t-k}^2-2\left(\frac{\sqrt{\frac{\alpha_t}{\alpha_{t-k}}} \boldsymbol{x}_t}{1-\frac{\alpha_t}{\alpha_{t-k}}}+\frac{\sqrt{\alpha_{t-k}} \boldsymbol{x}_0}{1-\alpha_{t-k}}\right) \boldsymbol{x}_{t-k}\right]\right\} \\
\propto & \exp\left\{-\frac{1}{2}\left(\frac{1-\alpha_t}{\left(1-\frac{\alpha_t}{\alpha_{t-k}}\right)\left(1-\alpha_{t-k}\right)}\right)\left[\boldsymbol{x}_{t-k}^2-2 \frac{\sqrt{\frac{\alpha_t}{\alpha_{t-k}}}\left(1-\alpha_{t-k}\right) \boldsymbol{x}_t+\sqrt{\alpha_{t-k}}\left(1-\frac{\alpha_t}{\alpha_{t-k}}\right) \boldsymbol{x}_0}{1-\alpha_t} \boldsymbol{x}_{t-k}\right]\right\} \\
= & \exp\left\{-\frac{\left(\boldsymbol{x}_{t-k}-\frac{\sqrt{\frac{\alpha_t}{\alpha_{t-k}}}\left(1-\alpha_{t-k}\right) \boldsymbol{x}_t+\sqrt{\alpha_{t-k}}\left(1-\frac{\alpha_t}{\alpha_{t-k}}\right) \boldsymbol{x}_0}{1-\alpha_t}\right)^2}{2\left(\frac{\left(1-\frac{\alpha_t}{\alpha_{t-k}}\right)\left(1-\alpha_{t-k}\right)}{1-\alpha_t}\right)}\right\} \\
\propto & \mathcal{N}\left(\boldsymbol{x}_{t-k} ; \underbrace{\frac{\sqrt{\frac{\alpha_t}{\alpha_{t-k}}}\left(1-\alpha_{t-k}\right) \boldsymbol{x}_t+\sqrt{\alpha_{t-k}}\left(1-\frac{\alpha_t}{\alpha_{t-k}}\right) \boldsymbol{x}_0}{1-\alpha_t}}_{\mu_q\left(\boldsymbol{x}_t, \boldsymbol{x}_0\right)}, \underbrace{\frac{\left(1-\frac{\alpha_t}{\alpha_{t-k}}\right)\left(1-\alpha_{t-k}\right)}{1-\alpha_t}}_{\boldsymbol{\Sigma}_q(t)} \mathbf{I}\right) .
\end{aligned}
\tag{9}
$$

For DDIM, we assume skip transitions are Gaussian distribution with unknown parameters:

$$
p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \kappa_t \boldsymbol{x}_t + \lambda_t \boldsymbol{x}_0, \sigma_t^2 \boldsymbol{I}). \tag{10}
$$

Using forward equations, Marginal consistency Eq. (4) gives the constraints:

$$
\lambda_{t,t-k} + \kappa_{t,t-k}\sqrt{\alpha_t} = \sqrt{\alpha_{t-k}}, \quad \kappa_{t,t-k}^2(1-\alpha_t) + \sigma_{t-k+1}^2 = 1-\alpha_{t-k}. \tag{11}
$$

Solving, we obtain:

$$
\kappa_{t,t-k} = \frac{\sqrt{1-\alpha_{t-k}-\sigma^2}}{\sqrt{1-\alpha_t}}, \quad \lambda_{t,t-k} = \sqrt{\alpha_{t-k}} - \frac{\sqrt{1-\alpha_{t-k}-\sigma^2}}{\sqrt{1-\alpha_t}}\sqrt{\alpha_t}. \tag{12}
$$

Substituting back yields

$$\begin{aligned} & P(x_{t-1}|x_t, x_0) \\ =& \mathcal{N}(x_{t-1}; (\sqrt{\alpha_{t-1}} - \frac{\sqrt{1-\alpha_{t-1}-\sigma^2}}{\sqrt{1-\alpha_t}}\sqrt{\alpha_t})x_0 + (\frac{\sqrt{1-\alpha_{t-1}-\sigma^2}}{\sqrt{1-\alpha_t}})x_t, \sigma_t^2) \\ =& \mathcal{N}(x_{t-1}; \sqrt{\alpha_{t-1}}x_0 + \sqrt{1-\alpha_{t-1}-\sigma^2}\epsilon_t, \sigma_t^2), \end{aligned} \tag{13}$$

and

$$x_{t-k} = \sqrt{\alpha_{t-k}} \underbrace{\left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}}\right)}_{\text{predicted } x_0} + \sqrt{1-\alpha_{t-k}-\sigma_t^2} \cdot \epsilon_\theta(x_t, t) + \sigma_t z_t. \tag{14}$$

8. Algorithm for Conservative Parallelization

Algorithm 2 Parallelization (Conservative Version)

Require: sampled noise x_T, total steps T, block size k
Ensure: generated image x_0

$t \leftarrow T$
repeat
  $\varepsilon_t \leftarrow \varepsilon_\theta(x_t, t)$
  for $i = 1$ to k **concurrently do** ▷ the following k noise prediction can run in parallel
    sample $z \sim \mathcal{N}(0, 1)$
    $x_{t-i} \leftarrow \sqrt{\alpha_{t-i}} \cdot \frac{x_t - \sqrt{1-\alpha_t} \cdot \varepsilon_t}{\sqrt{\alpha_t}} + \sqrt{1-\alpha_{t-i}-\sigma_{t,i}^2} \cdot \varepsilon_t + \sigma_{t,i} \cdot z$
    $\varepsilon_{t-i} \leftarrow \varepsilon_\theta(x_{t-i}, t-i)$ ▷ computation bottleneck
  end for
  for $i = 2$ to $k+1$ **do**
    sample $z \sim \mathcal{N}(0, 1)$
    $x_{t-i} \leftarrow \sqrt{\alpha_{t-i}} \cdot \frac{x_{t-i+1} - \sqrt{1-\alpha_{t-i+1}} \cdot \varepsilon_{t-i+1}}{\sqrt{\alpha_{t-i+1}}} + \sqrt{1-\alpha_{t-i}-\sigma_{t-i+1}^2} \cdot \varepsilon_{t-i+1} + \sigma_{t-i+1} \cdot z$
  end for
  $t \leftarrow t - k - 1$
until $t \leq 0$
return x_0

9. Additional Experiments

9.1. Quatitative Results on Other Models

We provide additional quantitative results on SD1.5 and SD2.1-base in Table 3. The overall trend is consistent with the main-text observations: DRiffusion continues to offer substantial wall-clock acceleration while maintaining quality comparable to the original sampler. On SD1.5, the method achieves up to 3.4× speedup, with FID, CLIP, Pick, and HPSv2.1 all remaining close to the original setting. A similar pattern is observed on SD2.1-base, where the speedup reaches 3.5×. These additional results further confirm that the favorable efficiency–quality trade-off of DRiffusion generalizes across different diffusion backbones.

9.2. Comparison with Distillation Method

While parallelization and distillation methods each offer distinct benefits, we highlight the compelling advantages of our approach over distillation. First and foremost, DRiffusion is strictly training-free. Unlike distillation, which necessitates costly and model-specific training, our method serves as a plug-and-play algorithmic acceleration universally applicable to any pre-trained model. Furthermore, our approach excels in preserving both alignment and generation diversity. Distillation techniques, such as LCM [23], often suffer from a noticeable distribution shift from the teacher model, leading to degraded

Table 3. Quantitative results of DRiffusion under SD1.5 and SD2.1-base on the MS-COCO dataset.

Model	Mode	#Devices	Latency (s)	Speed Up	FID ↓	CLIP ↑	Pick ↑	HPSv2.1 ↑
SD1.5	original	1	2.68	–	25.91	26.49	21.49	26.07
	conservative	2	2.07	1.3×	25.95	26.60	21.51	25.98
		3	1.43	1.9×	25.67	26.60	21.51	25.95
		4	1.10	2.4×	25.57	26.61	21.50	25.92
	aggressive	2	1.41	1.9×	26.04	26.55	21.50	25.94
		3	1.00	2.7×	25.48	26.58	21.49	25.90
		4	0.80	3.4×	25.36	26.55	21.45	25.81
SD2.1 base	original	1	2.48	–	25.69	26.19	21.81	27.14
	conservative	2	1.74	1.4×	25.75	26.26	21.83	27.04
		3	1.32	1.9×	25.67	26.27	21.82	27.02
		4	1.04	2.4×	25.56	26.26	21.81	26.99
	aggressive	2	1.35	1.8×	25.45	26.26	21.82	27.02
		3	0.92	2.7×	25.46	26.24	21.80	26.96
		4	0.72	3.5×	25.14	26.21	21.76	26.88

diversity. In contrast, DRiffusion faithfully retains the original model's behavior. As illustrated in Figure 6, while LCM deviates from the teacher's sampling trajectory and yields less diverse outputs, our method remains strictly aligned. Therefore, DRiffusion presents a more accessible acceleration paradigm, delivering substantial speedups without compromising the intrinsic quality or diversity of the original diffusion model.

10. Future Works

On the engineering side, an immediate direction is to extend DRiffusion to other generative modalities, such as video, policy generation, and audio. In these settings, diffusion models can serve as powerful and efficient generators of high-quality data for downstream tasks [39].

On the algorithmic side, future work could further exploit the temporal flexibility of diffusion. Although DRiffusion is studied here in the multi-device setting, it may also accelerate single-device sampling by enabling larger batch sizes. Another important direction is to establish theoretical guarantees that bound the deviation of the draft trajectory from the original one, potentially enabling speculative-decoding-like methods [16] for diffusion sampling.

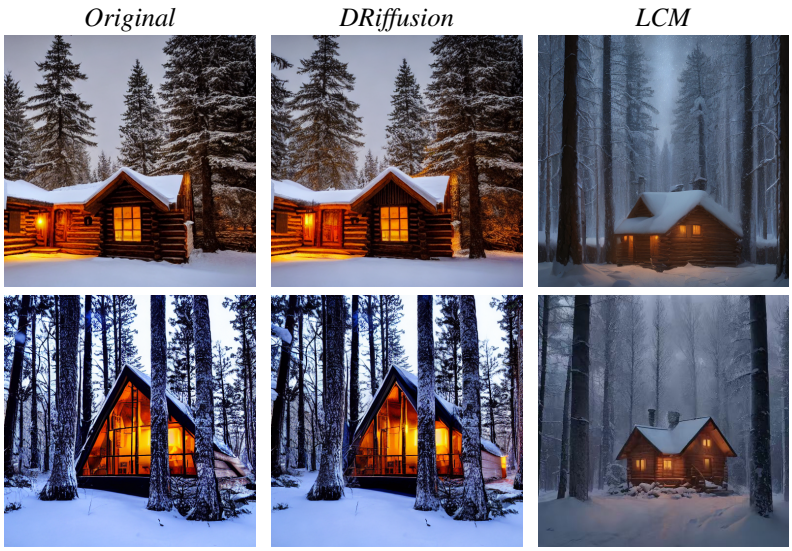

Figure 6. Comparison with LCM. We apply DRiffusion to LCM's teacher. *A cozy cabin in a snowy forest at night, glowing windows.*