

ParaFlow: Parallel Sampling for Flow Matching Models

Jianrong Lu¹, Bangwei Li¹, Haomin Zhang¹ , Zhuoya Gu¹ , Yongqing Lu²,
Jianhai Chen¹ , and Qinming He¹ 

¹ Zhejiang University, Hangzhou, China

jrong.alvin@gmail.com

{chenjh919, hqm}@zju.edu.cn

² Guizhou Normal University, Guiyang, China

A Use of LLM

During the preparation of this work, we used Large Language Models (LLMs) to assist with the writing process. The primary uses included polishing and improving the fluency of the text, generating preliminary drafts of proofs, and assisting in the creation and formatting of tables. After using these tools, the author(s) reviewed and edited the content extensively. We take full responsibility for the entire content of this publication, including the ideas, proofs, and presentations ultimately contained in the final manuscript.

B Proof of Proposition 1

We first show that the triangular nonlinear equation (TNE) system in Eq. (4) has a unique root. Suppose there exist two distinct solutions $\{A_0, \dots, A_N\}$ and $\{B_0, \dots, B_N\}$. By construction, for any $i \in \{1, \dots, N\}$ these solutions satisfy:

$$\begin{cases} A_i = F_{i-1}(A_0, \dots, A_{i-1}) \\ B_i = F_{i-1}(B_0, \dots, B_{i-1}). \end{cases} \quad (1)$$

By induction, assume $A_j = B_j$ for all $0 \leq j \leq i-1$. Then,

$$A_i = F_{i-1}(A_0, \dots, A_{i-1}) = F_{i-1}(B_0, \dots, B_{i-1}) = B_i, \quad (2)$$

which implies $A_i = B_i$. Therefore, all components of the two solutions must coincide, and the root is unique.

Next, we show that this unique root coincides with the autoregressive trajectory. From Eq. (3), the autoregressive update satisfies

$$x_{t_i} = x_{t_0} + \sum_{j=0}^{i-1} h_j v(x_{t_j}, t_j, \theta). \quad (3)$$

Meanwhile, the definition of F_{i-1} in Eq. (5) gives

$$\hat{x}_{t_i} = F_{i-1}(\hat{x}_{t_0}, \dots, \hat{x}_{t_{i-1}}) = \hat{x}_{t_0} + \sum_{j=0}^{i-1} h_j v(\hat{x}_{t_j}, t_j, \theta). \quad (4)$$

By induction on i , since $\hat{x}_{t_0} = x_{t_0}$, it follows that $\hat{x}_{t_i} = x_{t_i}$ for all $i \in \{0, \dots, N\}$. Thus, the TNE system admits a unique solution identical to the autoregressive trajectory.

C Proof of Proposition 2

We analyze the fixed-point iteration defined in Eq. (6):

$$\hat{x}_{t_0}^{(k+1)} = x_{t_0}, \quad \hat{x}_{t_i}^{(k+1)} = F_{i-1}(\hat{x}_{t_0}^{(k)}, \dots, \hat{x}_{t_{i-1}}^{(k)}), \quad i \in \{1, \dots, N\}. \quad (5)$$

We prove by induction that after k iterations, $\hat{x}_{t_j}^{(k)} = x_{t_j}$ for all $j \leq k$.

Base case ($k = 1$). By definition, $\hat{x}_{t_0}^{(1)} = x_{t_0}$. Moreover,

$$\hat{x}_{t_1}^{(1)} = F_0(\hat{x}_{t_0}^{(0)}) = F_0(x_{t_0}) = x_{t_1}. \quad (6)$$

Thus, indices 0 and 1 are exact after the first iteration.

Inductive step. Suppose $\hat{x}_{t_j}^{(k)} = x_{t_j}$ for all $j \leq k$. At the $(k+1)$ -th iteration,

$$\hat{x}_{t_{k+1}}^{(k+1)} = F_k(\hat{x}_{t_0}^{(k)}, \dots, \hat{x}_{t_k}^{(k)}) = F_k(x_{t_0}, \dots, x_{t_k}) = x_{t_{k+1}}, \quad (7)$$

Hence, the $(k+1)$ -th variable becomes exact, while all previously correct variables remain unchanged, since their updates depend only on values already exact.

Conclusion. After N iterations, $\hat{x}_{t_j}^{(N)} = x_{t_j}$ for all $j = 0, \dots, N$, establishing exact convergence in at most N steps.

D Proof of Proposition 3

Assumptions: We assume the velocity field $v(x, t)$ is smooth and satisfies the following conditions for constants $L_x, M_2 > 0$:

1. **Lipschitz continuity in space:** $\|v(x, t) - v(y, t)\| \leq L_x \|x - y\|$ for all $x, y \in \mathbb{R}^d$.
2. **Bounded second-order time-derivatives:** $\left\| \frac{\partial^2 v(x, t)}{\partial t^2} \right\| \leq M_2$ for all $x \in \mathbb{R}^d$.

Proof. Let $e_i := \|x_{t_i}^{approx} - x_{t_i}^{true}\|$ be the spatial error at timestep t_i . The forward Euler updates for the true and approximate trajectories are given by:

$$x_{t_{i+1}}^{true} = x_{t_i}^{true} + \Delta t \cdot v(x_{t_i}^{true}, t_i) \quad (8)$$

$$x_{t_{i+1}}^{approx} = x_{t_i}^{approx} + \Delta t \cdot \tilde{v}_i \quad (9)$$

where $\tilde{v}_i$ is the (potentially approximated) velocity used at step i .

By subtracting the two update equations and applying the triangle inequality, we derive the error recurrence:

$$e_{i+1} = \|x_{t_i}^{approx} - x_{t_i}^{true} + \Delta t(\tilde{v}_i - v(x_{t_i}^{true}, t_i))\| \leq e_i + \Delta t \cdot \|\tilde{v}_i - v(x_{t_i}^{true}, t_i)\|. \quad (10)$$

We bound the velocity mismatch term by splitting it into the approximation error and the propagated error:

$$\|\tilde{v}_i - v(x_{t_i}^{true}, t_i)\| \leq \|\tilde{v}_i - v(x_{t_i}^{approx}, t_i)\| + \|v(x_{t_i}^{approx}, t_i) - v(x_{t_i}^{true}, t_i)\|. \quad (11)$$

The propagated error is bounded by the spatial Lipschitz condition:

$$\|v(x_{t_i}^{approx}, t_i) - v(x_{t_i}^{true}, t_i)\| \leq L_x e_i. \quad (12)$$

For the approximation error, we analyze the scenario where an approximation occurs. Since an estimation happens at each parallel iteration, it occurs a total of K times throughout the sampling process. If no approximation occurs at step i , this error is exactly 0.

Using the remainder theorem for Newton polynomial interpolation of degree 1, the extrapolation error at time t_i using reference points t_c and t_p is bounded by:

$$\|\tilde{v}_i - v(x_{t_i}^{approx}, t_i)\| \leq \frac{M_2}{2} |(t_i - t_c)(t_i - t_p)|. \quad (13)$$

Given a uniform step size $\Delta t = 1/N$ and a maximum sliding window size of p , the distance between the target step and the historical references is at most $p\Delta t$. Therefore, we can bound the temporal differences as $|t_i - t_c| \leq p\Delta t$ and $|t_i - t_p| \leq (p+1)\Delta t$.

This yields the upper bound for the velocity mismatch:

$$\|\tilde{v}_i - v(x_{t_i}^{approx}, t_i)\| \leq \frac{M_2 p(p+1)}{2} (\Delta t)^2. \quad (14)$$

Combining these bounds, the error recurrence becomes:

$$e_{i+1} \leq (1 + L_x \Delta t) e_i + \mathbb{I}_{\{\text{est. at } i\}} \cdot \frac{M_2 p(p+1)}{2} (\Delta t)^3. \quad (15)$$

Unrolling this recurrence from the initial condition $e_0 = 0$ yields:

$$e_N \leq \sum_{i=0}^{N-1} \mathbb{I}_{\{\text{est. at } i\}} \cdot \frac{M_2 p(p+1)}{2} (\Delta t)^3 \cdot (1 + L_x \Delta t)^{N-1-i}. \quad (16)$$

Using the inequality $(1+x) \leq e^x$, we bound the exponential term as $(1 + L_x \Delta t)^{N-1-i} \leq e^{L_x N \Delta t} = e^{L_x}$. Substituting this back, we get a summation over the exactly K estimated segments:

$$e_N \leq \sum_{k=1}^K \frac{M_2 p(p+1)}{2} (\Delta t)^3 \cdot e^{L_x} = K \cdot \frac{M_2 p(p+1) e^{L_x}}{2} (\Delta t)^3. \quad (17)$$

Table 1: Comparison with representative parallel sampling methods on the Flux model using 50 Euler sampling steps on a single GPU. Existing methods incur substantial NFE overhead, resulting in slower inference than the sequential Euler baseline, whereas ParaFlow achieves significant acceleration while maintaining comparable generation quality.

Method (Flux, 50 steps, 1 GPU)	CLIP $\uparrow$	FID $\downarrow$	Speedup $\uparrow$
Euler baseline	32.38	61.36	1.0 $\times$
ParaDiGMS	32.44	61.34	0.4 $\times$
ParaSolver	32.39	61.36	0.6 $\times$
CHORDS	32.41	61.35	0.8 $\times$
ParaFlow (Ours)	32.43	61.34	2.8$\times$

Since $\Delta t = 1/N$, the bound becomes:

$$e_N \leq \left(\frac{K M_2 p(p+1) e^{L_x}}{2} \right) \frac{1}{N^3} \quad (18)$$

which conclusively proves $e_N = \mathcal{O}\left(\frac{K}{N^3}\right)$.

E Additional Comparison with Existing Parallel Samplers

We further compare ParaFlow with representative parallel sampling methods, including ParaDiGMS, ParaSolver, and CHORDS, under the same experimental setting on the Flux model with 50 Euler sampling steps using a single GPU. Although these methods introduce step-level parallelism, they typically require substantially more function evaluations (NFEs) to ensure convergence, resulting in considerable computational overhead. Consequently, the increased computation offsets the benefit of parallel execution, leading to slower wall-clock inference than the standard sequential Euler solver.

In contrast, ParaFlow leverages the proposed velocity approximation strategy to significantly reduce the number of required model evaluations while preserving the convergence behavior of the original autoregressive sampler. As shown in Table 1, ParaFlow achieves a 2.8 $\times$ wall-clock speedup over the Euler baseline while maintaining essentially identical generation quality in terms of both CLIP Score and FID. These results further demonstrate that minimizing NFE overhead is crucial for realizing practical parallel sampling on commodity single-GPU platforms.

F Additional Qualitative Results



(a) A peaceful river surrounded by tall trees and lush greenery on both sides. (b) The glowing ember of a dying fire, surrounded by the charred remains of last night's logs and a few wispy strands of smoke curling upwards. (c) Women in swimwear performing synchronized swimming routine. (d) A stunning mountain range at sunset, with snow-capped peaks fading into a brilliant orange and pink sky that stretches endlessly in every direction.



(e) Frozen lake with frosty trees reflected in the water (f) A stunning portrait of a wise old woodworker standing in front of a wooden workbench, surrounded by tools and half-finished projects. (g) LED-lit cityscape at night (h) Old, creepy carnival at midnight

Fig. 1: Qualitative results generated by Euler + ParaFlow with different prompts on the FLUX.1-dev models at a resolution of 1024×1024 .



(a) Envision a serene coastal scene featuring a rustic lighthouse situated among towering sand dunes and vast expanses of sandy beach, as the sun rises over the horizon. (b) An intricately carved Icelandic turf roofed cabin standing proudly amidst a windswept landscape of black sand beaches and jagged rock formations. (c) Show a dramatic split-second moment of a thunderstorm unleashing its fury on a deserted highway, with lightning flashing across the darkening sky and debris scattered everywhere. (d) Depict Captain America standing on a rooftop, looking out over a cityscape at sunset, with his eyes narrowed in focus as he surveys his surroundings.



(e) A pod of dolphins swimming through coral reef (f) Desert landscape under stars (g) A massive thunderstorm unfolding over the Great Sandy Desert, with towering cumulus clouds stretching across the sky. (h) A vibrant pomegranate fruit sitting alone on a rustic wooden table, surrounded by lush greenery and autumn leaves.

Fig. 2: Qualitative results generated by Euler + ParaFlow with different prompts on the stable-diffusion-3-medium models at a resolution of 1024×1024 .