ICLR

FURIOSA

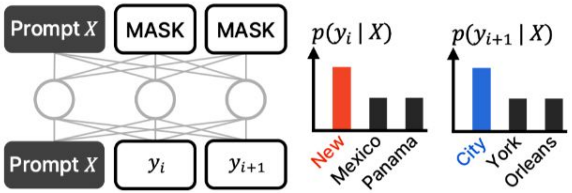
ParallelBench: Understanding the Trade-offs of Parallel Decoding in Diffusion LLMs

Wonjun Kang*, Kevin Galim*, Seunghyuk Oh*, Minjae Lee, Yuchen Zeng, Shuibai Zhang, Coleman Hooper, Yuezhou Hu, Hyung Il Koo, Nam Ik Cho, Kangwook Lee

Issue Limitations of Parallel Decoding

Q. Pick a random city for travel: New York, New Orleans, Mexico City, or Panama City?

	y_i	y_{i+1}	Joint
One-by-One	$p(y_i \mid X)$	$p(y_{i+1} \mid X, y_i)$	$p(y_i, y_{i+1} \mid X)$
Parallel	$p(y_i \mid X)$	$p(y_{i+1} \mid X)$	$p(y_i \mid X) \cdot p(y_{i+1} \mid X)$



A. New City

Parallel decoding ignores token dependencies

Analysis A Case Study on List Operations

Q. Replace a random item in A, B, C with F.

A = 66%	B = 66%	C = 66%
F = 33%	F = 33%	F = 33%

$$\text{Acc.}* = 3\times\left(\frac{1}{3}\times\frac{2}{3}\times\frac{2}{3}\right) = 44.4\%$$

*Assuming that all tokens are unmasked in parallel

Q. Shuffle the following items: A, B, C.

A = 33%	A = 33%	A = 33%
B = 33%	B = 33%	B = 33%
C = 33%	C = 33%	C = 33%

$$\text{Acc.}* = \frac{3}{3}\times\frac{2}{3}\times\frac{1}{3} = 22.2\%$$

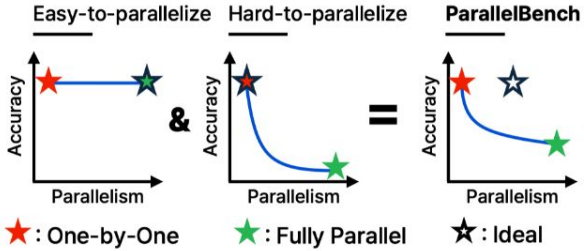
Quality inevitably drops under parallel decoding

Benchmark ParallelBench

Consists of 17 tasks across 3 categories

	Waiting Line	Text Writing	Puzzles
Metric	Accuracy	Grammar Score	Accuracy
# Tasks	10	5	2

Covers a wide-range of difficulty levels

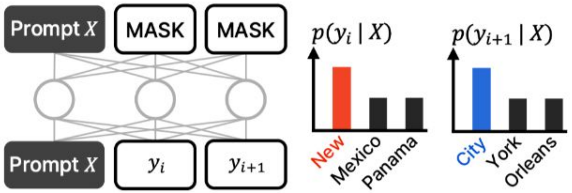


The first benchmark to assess the trade-offs of parallel decoding in dLLMs

Issue **Limitations of Parallel Decoding**

Q. Pick a random city for travel: New York, New Orleans, Mexico City, or Panama City?

	y_i	y_{i+1}	Joint
One-by-One	$p(y_i\|X)$	$p(y_{i+1}\|X, y_i)$	$p(y_i, y_{i+1}\|X)$
Parallel	$p(y_i\|X)$	$p(y_{i+1}\|X)$	$p(y_i\|X) \cdot p(y_{i+1}\|X)$



A. New City

Parallel decoding ignores token dependencies

Conditional Total Correlation C(Y|X):

$$\mathcal{C}(Y|X) = -H_{\text{data}}(Y|X) + \sum_{y_i \in Y} H_{\text{data}}(y_i|X)$$

One-Step Generation[1]:

[1] On the Learning of Non-Autoregressive Transformers (ICML 2022)

$$\min_\theta \mathcal{D}_{\text{KL}}(P_{\text{data}}(Y|X) \parallel P_\theta(Y|X)) \geq \mathcal{C}(Y|X)$$

T-Step Parallel Decoding:

Theorem 1 (Lower Bound for T-step Parallel Decoding). *For a factorized generative model (e.g., dLLM) $P_\theta(Y|X)$ performing T-step parallel decoding, assume the target sequence Y is partitioned into T disjoint sets, $Y = S_1 \cup S_2 \cup \cdots \cup S_T$. At each step $i \in \{1, \ldots, T\}$, the model generates the tokens in set S_i in parallel, conditioned on X and all previously generated tokens $S_{<i} = S_1 \cup S_2 \cup \ldots \cup S_{i-1}$ ($S_{<1} = \emptyset$). The minimum achievable KL divergence for this model is lower-bounded by:*

$$\min_\theta \mathcal{D}_{KL}(P_{data}(Y|X) \parallel P_\theta(Y|X)) \geq \mathcal{L}_T\left(\{S_i\}_{i=1}^T\right) := \sum_{i=1}^T \mathbb{E}_{S_{<i} \sim P_{data}}[\mathcal{C}(S_i|X, S_{<i})] \quad (2)$$

Analysis **A Case Study on List Operations**

Q. Replace a random item in A, B, C with F.

A = 66%	B = 66%	C = 66%
F = 33%	F = 33%	F = 33%

$$\text{Acc.}* = 3\times\left(\frac{1}{3}\times\frac{2}{3}\times\frac{2}{3}\right) = 44.4\%$$

*Assuming that all tokens are unmasked in parallel

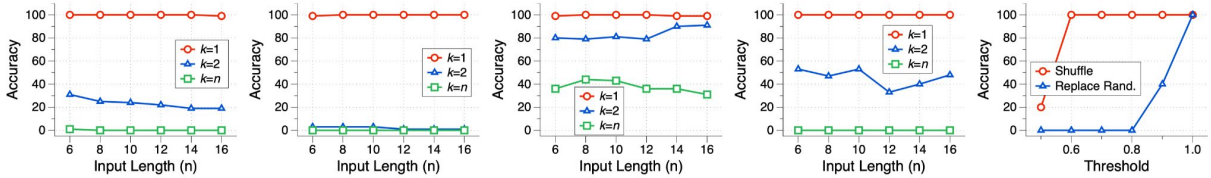
Q. Shuffle the following items: A, B, C.

A = 33%	A = 33%	A = 33%
B = 33%	B = 33%	B = 33%
C = 33%	C = 33%	C = 33%

$$\text{Acc.}* = \frac{3}{3}\times\frac{2}{3}\times\frac{1}{3} = 22.2\%$$

Quality inevitably drops under parallel decoding

Task / Analysis	**Data Distribution Perspective** (Section 4.1)		**Decoding Strategy Perspective** (Section 4.2)	
	$\mathcal{C}(Y\|X)$	$\lim_{n\to\infty}\mathcal{C}(Y\|X)$	Acc. (Greedy, Top-k)	Acc. (Greedy, Threshold)
Copy & Replace Index	0	0	1	1
Replace Random	$(n-1)[\log_2(n)-\log_2(n-1)]$	$\log_2(e)\approx 1.44$	0.5 if $k=2$; 0 if $k>2$	1 if $\gamma \geq (n-1)/n$; else 0
Shuffle	$n\log_2(n)-\log_2(n!)$	∞	Eq. (5)	1 if $\gamma > 0.5$



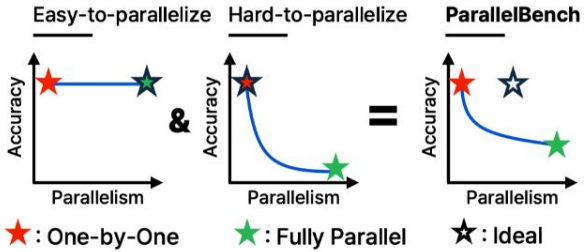
(a) Shuffle ($\tau = 1$) (b) Shuffle ($\tau = 0$) (c) Replace ($\tau = 1$) (d) Replace ($\tau = 0$) (e) Thres. ($n = 10$)

Benchmark **ParallelBench**

Consists of 17 tasks across 3 categories

	Waiting Line	Text Writing	Puzzles
Metric	Accuracy	Grammar Score	Accuracy
# Tasks	10	5	2

Covers a wide-range of difficulty levels



The first benchmark to assess the trade-offs of parallel decoding in dLLMs

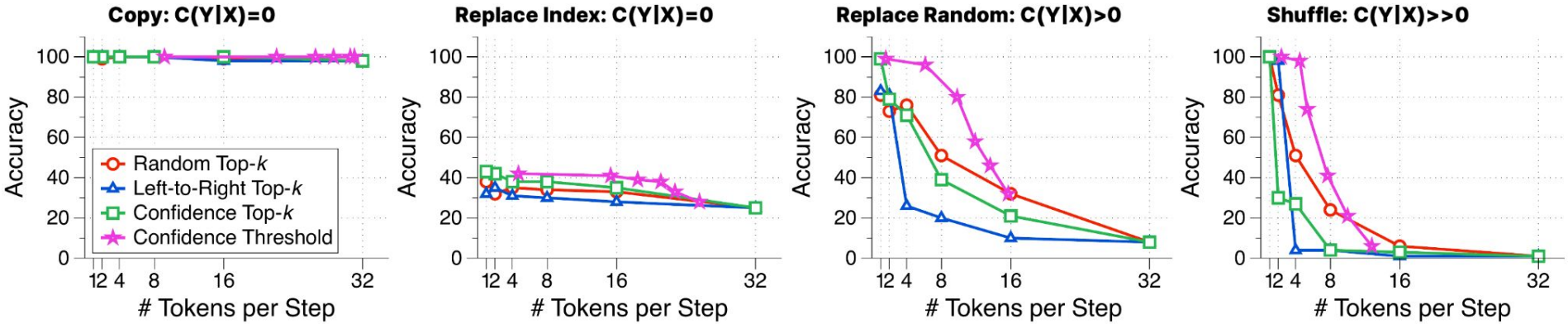
Shuffle

```
You are managing a waiting line at a customer service desk.
The waiting line should be randomly shuffled to ensure fair service distribution:

["Paul Payne", "Robert Riley", "Peter Stone"]

Please randomly shuffle the list and provide only the final list.
Ensure the sequence is different from the original.
```

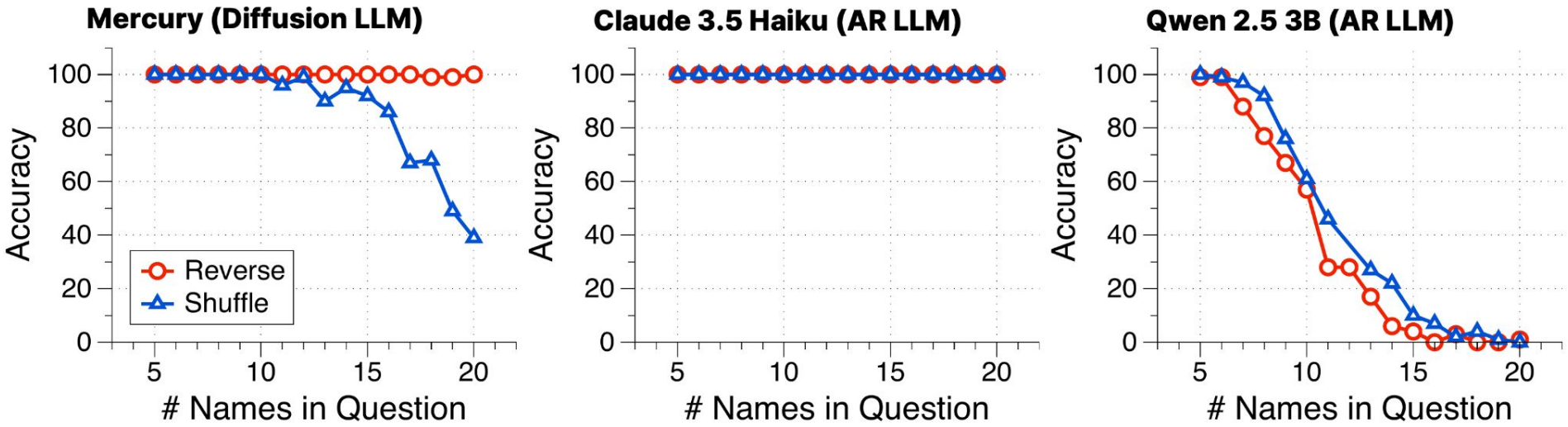
Words-to-Sentence (easy)

```
Construct a single, coherent sentence using the words dog, park, ball, and throw.
```

Benchmark results (LLaDA 1.5)



Diffusion LLM vs. Autoregressive LLM



7 Exploring Additional Techniques

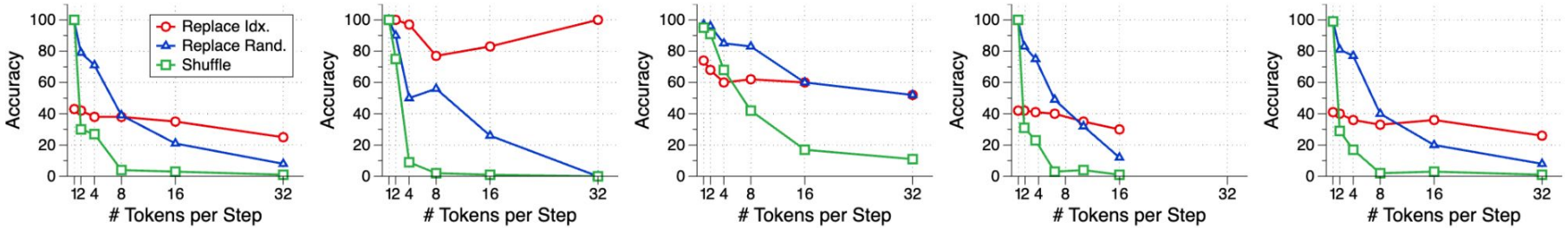
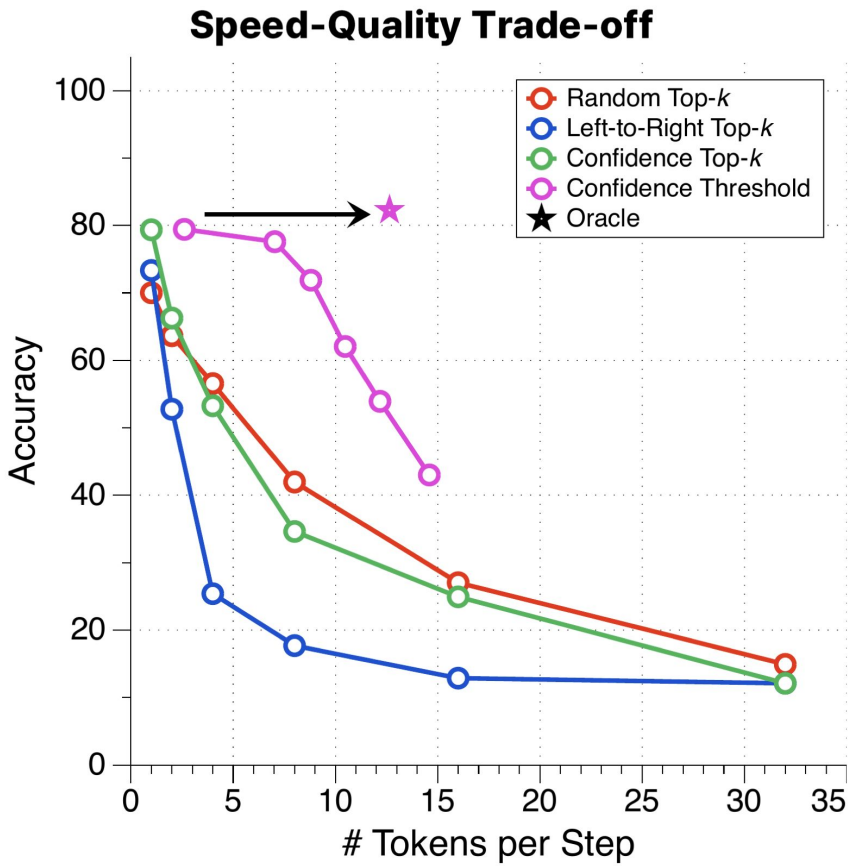


Figure 8: *Waiting Line* results using LLaDA 1.5 (*Random Top-k*) with various advanced techniques.

We further explore whether benchmark performance can be improved when various advanced techniques for dLLMs are applied. Details and full results are in Appendix E.



Takeaway. Static parallel decoding (e.g., top-k) can suffer severe quality degradation, and adaptive decoding strategies (e.g., threshold) still have significant room for improvement.

Thank You