DIVERSED: Relaxed Speculative Decoding via Dynamic Ensemble Verification

Ziyi Wang[1], Siva Rajesh Kasa[2], Ankith M S[2], Santhosh Kumar Kasa[2], Jiaru Zou[3],Sumit Negi[2], Ruqi Zhang[1], Nan Jiang[4], Qifan Song[1]
[1]Purdue University, [2]Amazon, [3]UIUC, [4]University of Texas – El Paso

Background & Motivating example

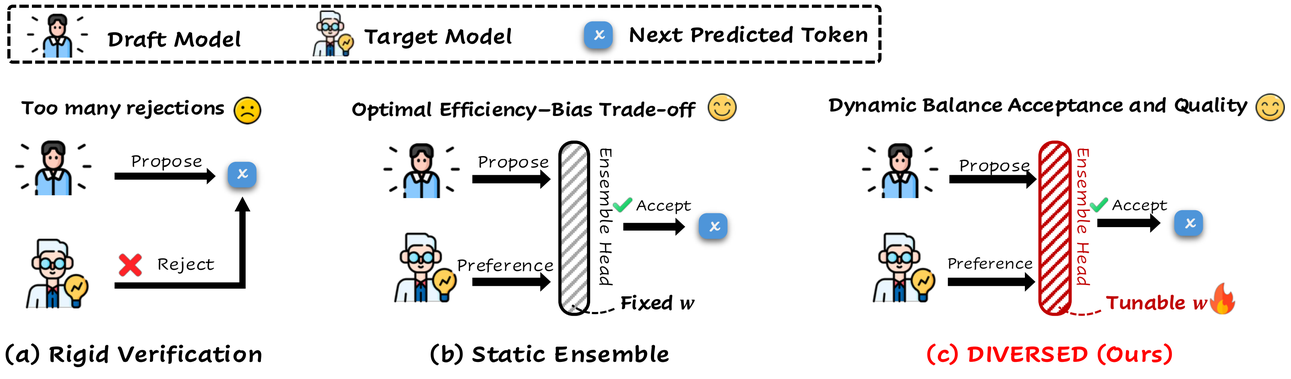
Task: Given a prompt $\mathbf{x}_0$, sample sequence $(\mathbf{x}_0, x_1, x_2 \ldots, x_n)$ from large target model $p_\theta(\cdot \,|\, x_0, \ldots, x_{t-1})$.
Procedure: using small proposal model $q_\theta(\cdot \,|\, x_0, \ldots, x_{t-1})$ to draw samples, Then the target model accept it with acceptance ratio b_t.

[Prompt] Each bird eats 12 beetles per day, ..., how many beetles are eaten each day?
Each jaguar eats 5 snakes per day, so 6 jaguars will eat 6 ... per day. Each bird eats 12 beetles day, and
[and] there ...✓ [So], each bird...✓
there are 90 birds, so in total they eat 12 * 90 = 1080 beetles per day.
[90] birds eat 12 * 90 = 1080 beetles per day. ✓

[Prompt] Alexis is applying for a new job ... How much did Alexis pay for the shoes?
Alexis spent $30 + ... + $18 = $143 on the items she has receipts for. She has $16 left from her budget, so
[$153] on the ... she spent $49. ×
she spent ... $184 in total. Therefore, she spent $184 − $143 = $41 on the shoes.
[on] the shoes ...×

Figure 3: Examples of speculative-decoding verification with accepted mismatches, illustrating not all draft–target mismatches are equally harmful. **Black** marks tokens both models agree on; **green** marks accepted mismatches that still yield the correct answer; **red** marks accepted mismatches that lead to an incorrect answer.

Our method: Dynamic Verification Relaxed Speculative Decoding



Compared with classic speculative decoding (shown in (a) rigid verification), our Diversed (c) achieves a higher acceptance rate, comparable accuracy, and lower wall-clock time.

Compared with static ensemble (b), Diversed (c) attains higher accuracy via a tunable ensemble weight that adapts to the task and the context.

Code Implementation is here!

Experiments

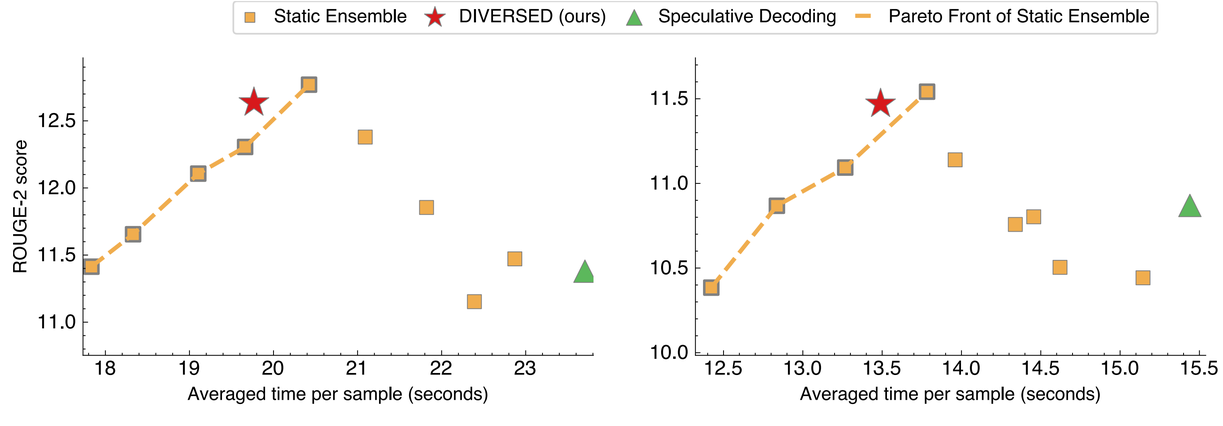


Figure 2. Our method improves upon the Pareto frontier achieved by static ensemble, attaining better trade-offs between inference time and generation quality.

Results are reported on the CNNDM dataset using the target/draft model pair:Qwen3-8B/Qwen3-0.6B and Gemma-3-12b-it/Gemma-3-4b-it, with temperature being 0.

Table 1. Benchmark acceptance rate and generation quality. Columns group results by task while panels (a)–(c) correspond to different Target/Draft model pairs. All experiments use temperature 1 and draft length N = 5.

Our method achieves a higher acceptance rate and also maintains comparable quality.

Method	GSM8K		CNNDM		XSum		MBPP	
	Accept Rate (↑)	Quality Accuracy (↑)	Accept Rate(↑)	Quality ROUGE-2 (↑)	Accept Rate(↑)	Quality ROUGE-2 (↑)	Accept Rate(↑)	Quality pass@1 (↑)
Autoregressive	NA	67%	NA	9.86	NA	7.03	NA	53%
SD	44.60%	67%	21.60%	9.46	20.44%	7.09	26.30%	53%
SD (Lossy)	59.81%	66%	38.86%	10.51	40.91%	7.96	66.75%	49%
SpecCascade	61.53%	67%	47.29%	11.74	43.95%	7.61	73.92%	52%
Static Ensemble	69.49%	66%	61.06%	11.46	51.58%	7.22	68.70%	52%
DIVERSED (ours)	**72.61%**	67%	**69.96%**	12.11	**70.53%**	7.23	**85.03%**	53%
(a) Target/Draft model pair is Llama-3.1-8B/Llama-3.2-1B.								
Autoregressive	NA	90%	NA	9.97	NA	4.90	NA	55%
SD	59.58%	91%	35.26%	9.85	15.60%	4.95	58.72%	55%
SD (lossy)	60.10%	86%	36.40%	10.53	19.56%	4.90	60.85%	53%
SpecCascade	58.65%	86%	38.13%	10.85	19.45%	4.97	71.27%	56%
Static Ensemble	67.58%	86%	41.07%	10.88	21.24%	4.98	76.11%	55%
DIVERSED (ours)	**76.48%**	88%	**46.59%**	10.97	**47.01%**	5.01	**81.46%**	59%
(b) Target/Draft model pair is Qwen3-8B/Qwen3-0.6B.								
Autoregressive	NA	93%	NA	9.01	NA	8.31	NA	68%
SD	84.15%	92%	40.39%	9.06	35.76%	8.27	83.25%	67%
SD (Lossy)	85.02%	90%	45.43%	10.62	39.48%	8.38	86.69%	65%
SpecCascade	84.43%	92%	51.44%	10.42	39.42%	8.27	83.76%	67%
Static Ensemble	87.62%	91%	54.48%	10.84	61.42%	7.90	86.46%	66%
DIVERSED (ours)	**90.70%**	92%	**66.90%**	10.86	**63.38%**	7.22	**90.23%**	67%
(c) Target/Draft model pair is Gemma-3-12B/Gemma-3-4B.								

Ziyi Wang conducted this work during an Amazon internship. Nan Jiang is supported by TACC CCR25054.
Contact Email: njiang@utep.edu.