ReSET: Accurate Latency-Critical NVFP4 Reasoning via Step-Aware Temperature Scaling

Sihwa Lee[1*], Janghwan Lee[1*], Donghoon Yoo[2], Jaegon Kim[2], Hanyul Ryu[2], Soojung Ryu[2], Jungwook Choi[1†]

[1]Hanyang University [2]Xenoscube Korean Inc. [†]Corresponding Author: choij@hanyang.ac.kr



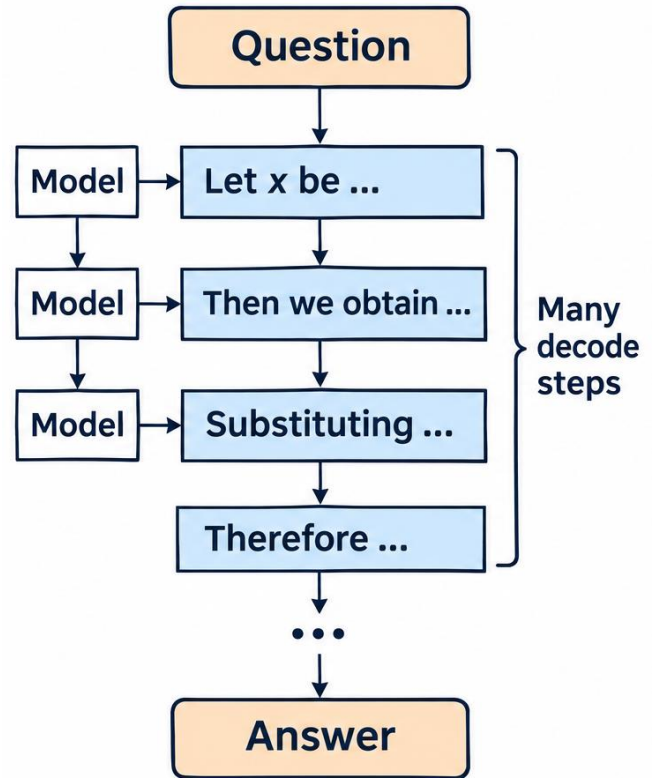




Efficient Reasoning with NVFP4

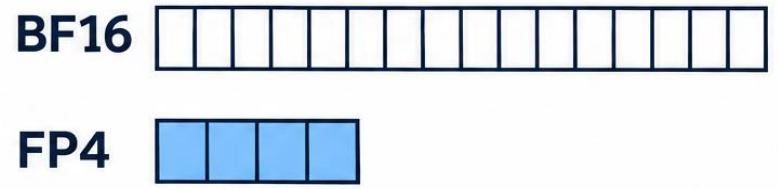
Long reasoning traces make efficient inference essential

Long Reasoning Traces

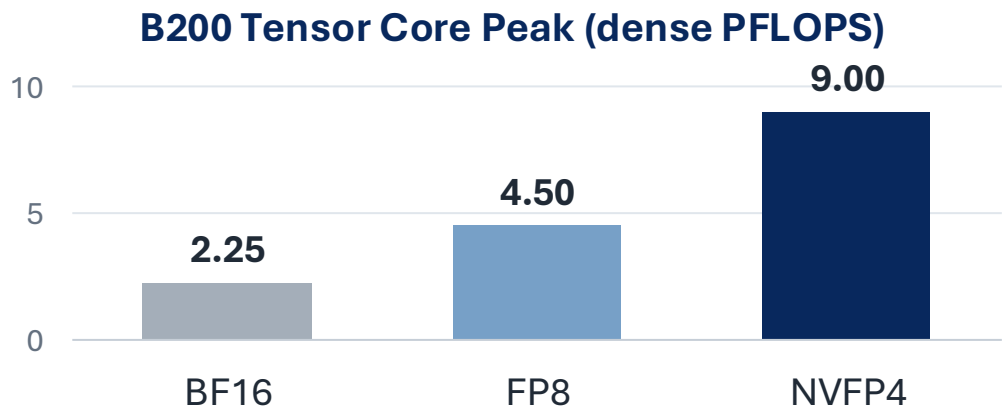


Repeated decoding increases cost

NVFP4 on Blackwell



Plus NVFP4 scaling metadata



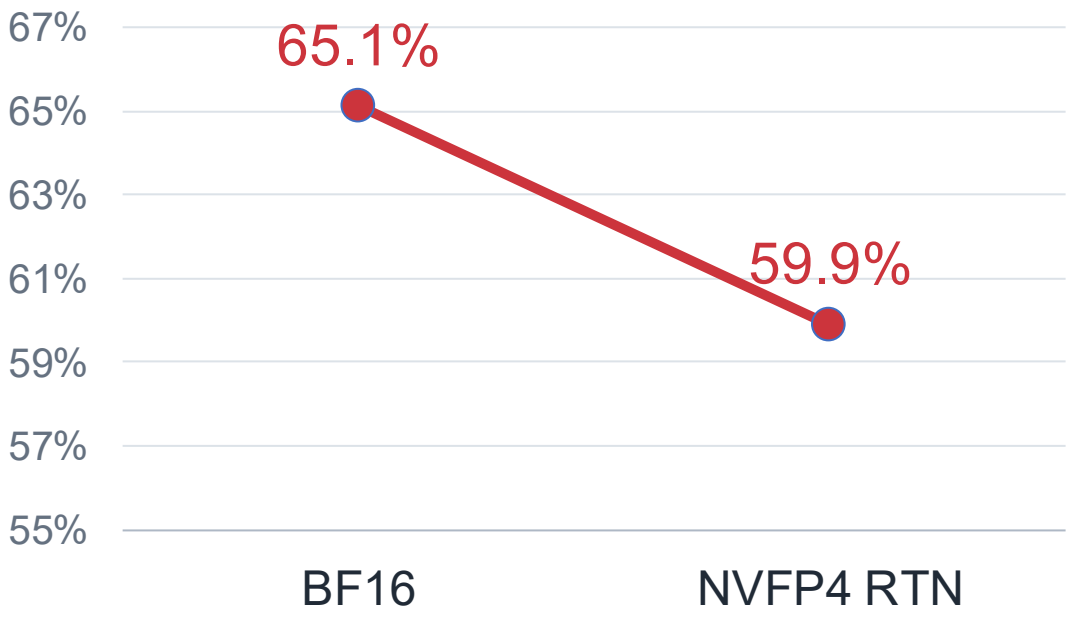
4× peak throughput over BF16

Source: NVIDIA Blackwell Datasheet, p. 8. Dense peak per B200 GPU.

Challenge 1: Accuracy Degradation under NVFP4

Quantizing the LLM to NVFP4 can substantially reduce reasoning accuracy

Five-model mean AIME-120 accuracy



−5.2 points

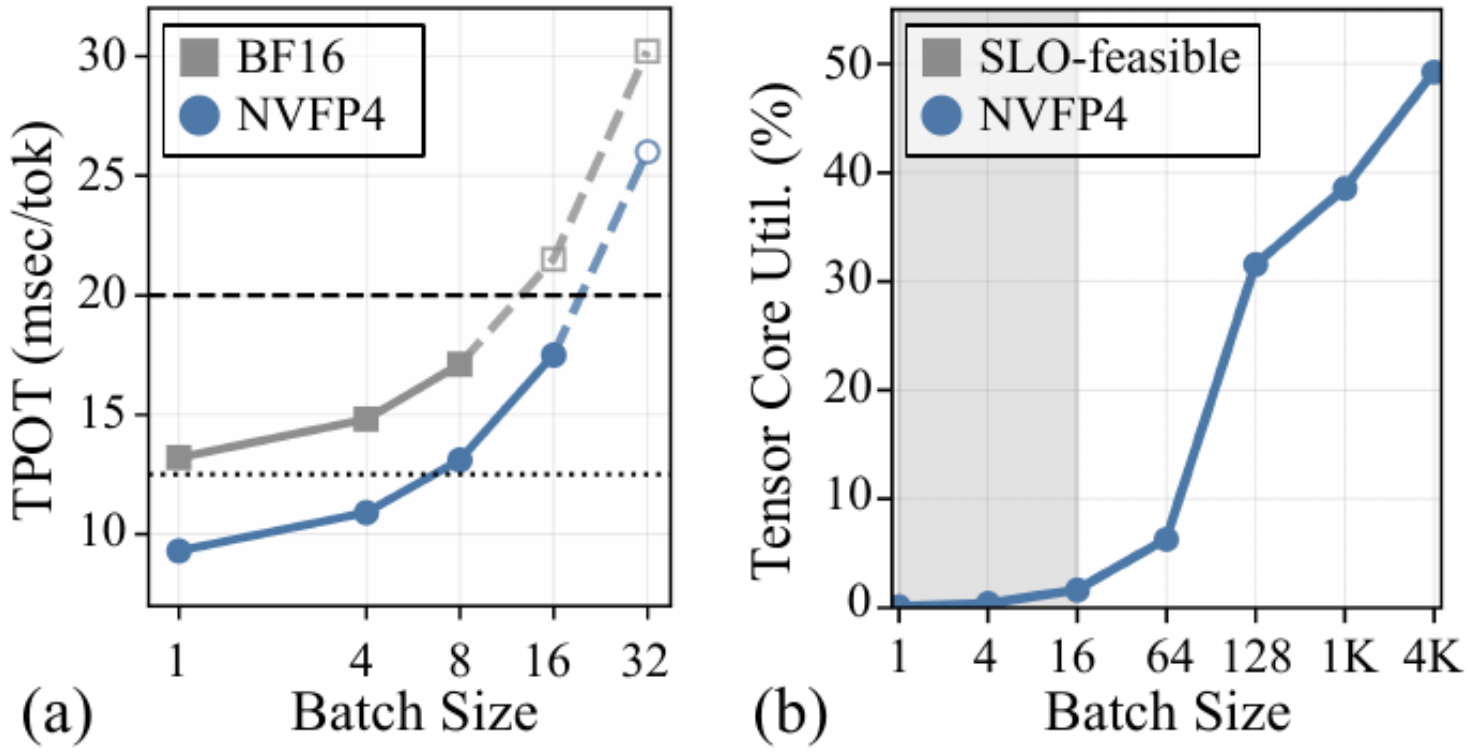
Model	BF16	NVFP4	Loss
R1-Qwen 7B	45.7	39.6	**−6.1**
R1-Qwen 14B	57.4	52.4	**−5.0**
Qwen3 8B	70.4	62.5	**−7.9**
Qwen3 14B	76.1	70.4	**−5.7**
Qwen3 32B	75.8	74.4	**−1.4**

Up to 7.9-point loss

AIME-120, eight seeds. Loss in percentage points.

Challenge 2: Inefficient Small-Batch Decoding

Small batches limit the benefit of high peak NVFP4 throughput



Long reasoning traces

Per-token latency sets the serving constraint

M ≤ 8

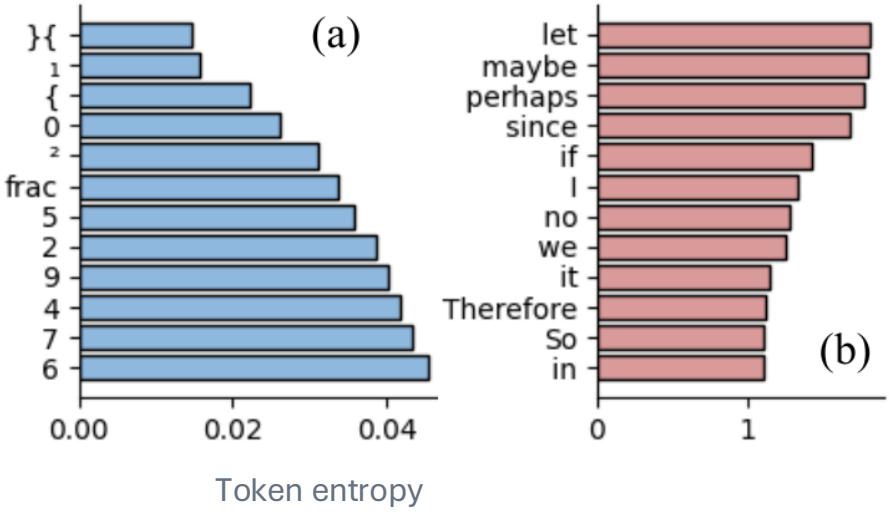
<1% utilization

The small-batch regime needs its own kernel

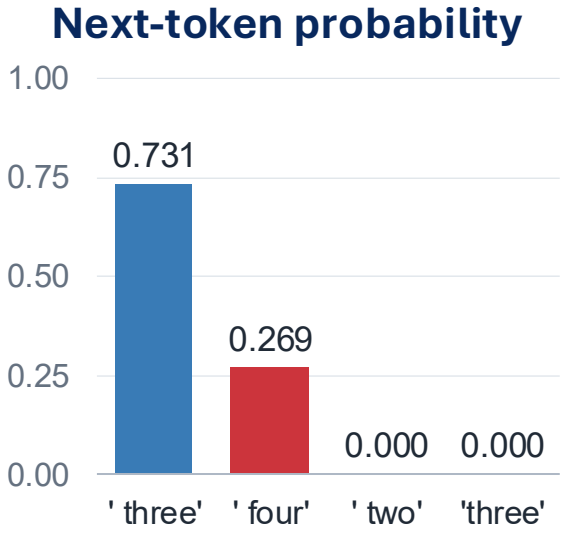
Single B200, Qwen3-32B, 32K context. M denotes the number of active decode tokens.

Sampling Errors at Low-Entropy Tokens

Symbolic tokens can remain top-ranked but still be sampled incorrectly



Low entropy: symbolic decisions
High entropy: branching choices



99.91%

BF16 top-1 stays top-1

1.54×

non-top-1 sampling probability

(c) **<BF16 Inference>** ... uses identity |A|+|B|+|C|+|D| = E1 + 2·E2 + 3·E3 + 4·E4\n 195+367+562+900 ... = 229 − x \n 2024 = (229 − x) + 874 + 702 + 4x \n 2024 = 1805 + 3x \n 3x = 219 \n x = **73**

(d) **<NVFP4 Inference>**... "+x +y +z −w = 900", ... owning exactly **four** things would be equal ... "exactly three = D + 2 of {A,B,C}" = 237 ... "exactly four = D + 3 of {A,B,C}" = 900 − 402 − 237 → x = **261**

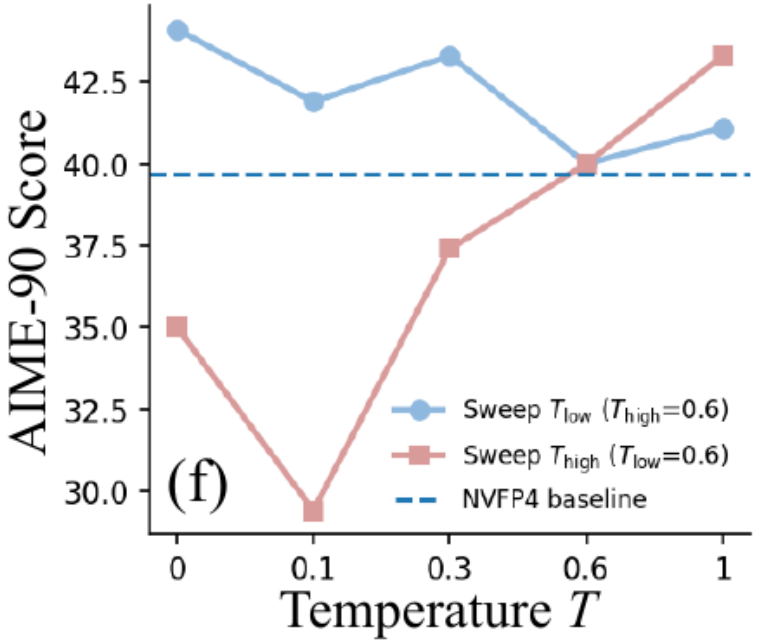
Example: R1-Qwen-7B. Matched-prefix analysis: Qwen3-32B, 46,375 low-entropy positions.

Temperature Scaling and Step Context

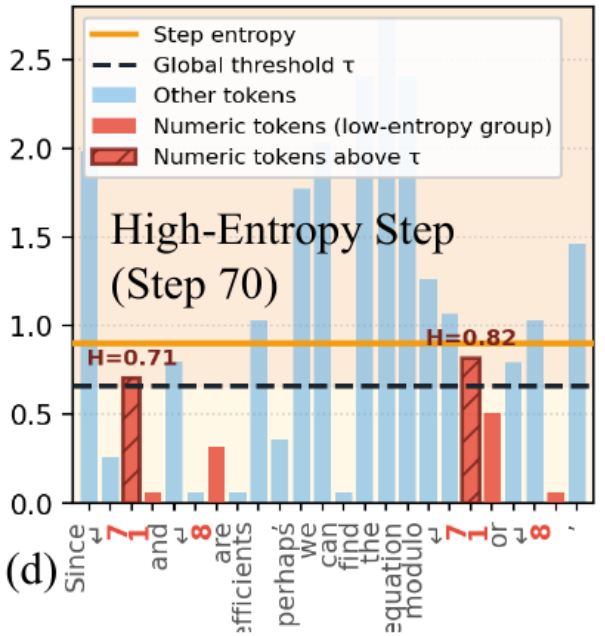
Lower temperature helps, but the threshold must reflect the reasoning step

Lower temperature at low entropy

Temperature sweep: R1-Qwen-7B, AIME-90



Token entropy in an uncertain step

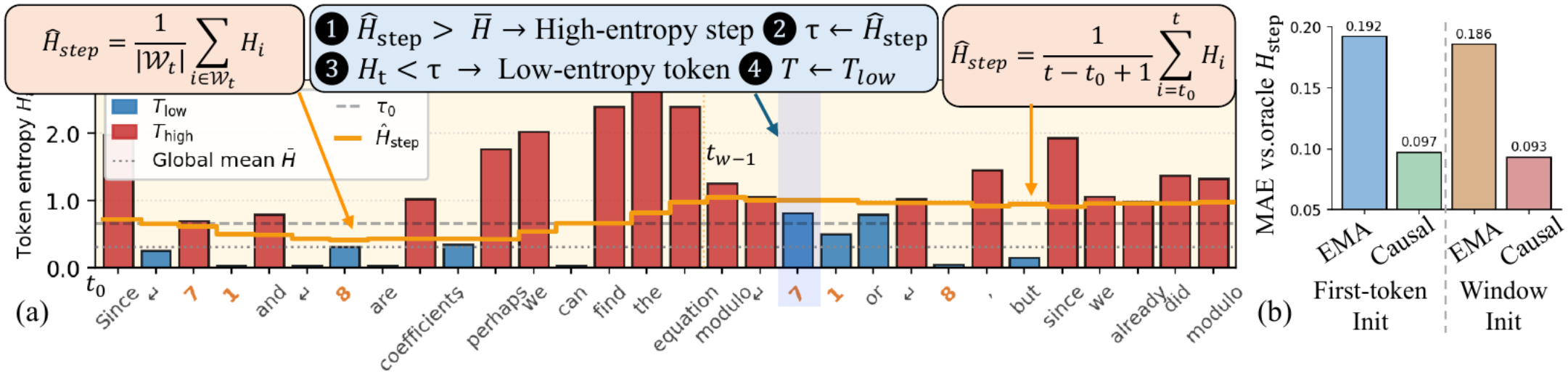


Lowering temperature at low-entropy positions improves NVFP4 accuracy

High-entropy steps can push symbolic tokens above the fixed threshold

ReSET: Step-Aware Temperature Scaling

ReSET combines a step-aware threshold with an online step-entropy estimate



① Step-aware threshold

Confident step: retain τ_0
Uncertain step: use estimated step entropy
Sharpen tokens below the threshold

② Online estimation

First 32 tokens of each step: window mean
Then: within-step running mean
No future tokens or model retraining

ReSET Results on Reasoning Benchmarks

Accuracy across five models and three benchmarks

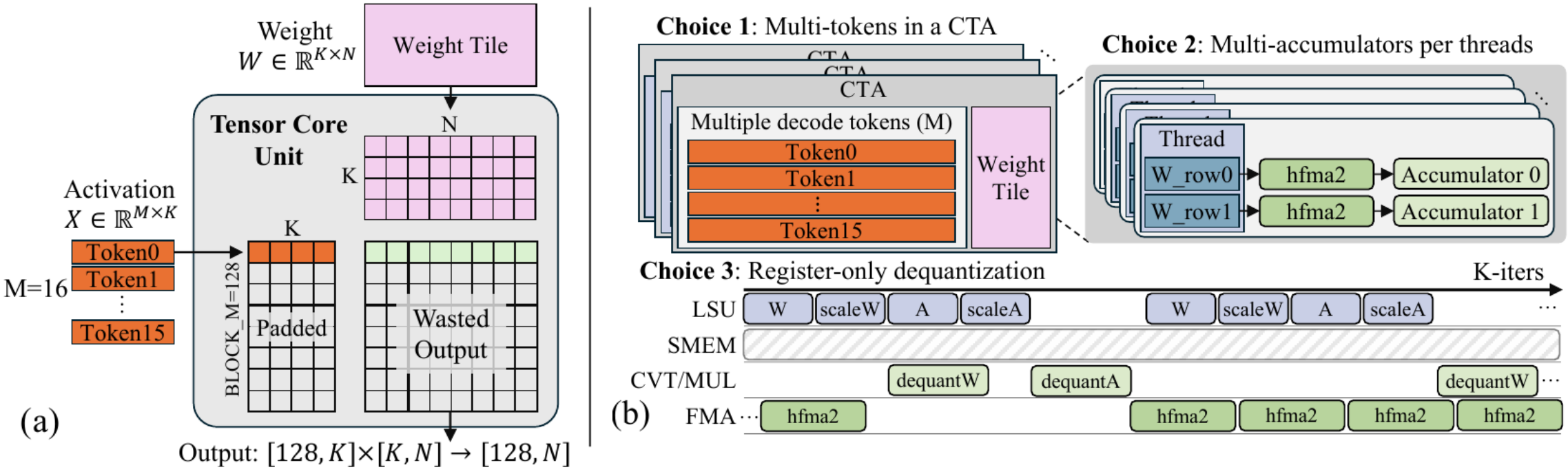
Task	Method	R1-Qwen 7B	R1-Qwen 14B	Qwen3 8B	Qwen3 14B	Qwen3 32B	Avg
AIME-120	BF16 Baseline	45.7	57.4	70.4	76.1	75.8	65.1
	RTN	39.6	52.4	62.5	70.4	74.4	59.9
	BRQ	41.4	49.8	53.8	66.9	73.0	57.0
	4/6	41.1	53.1	64.0	70.1	74.8	60.6
	MR-GPTQ	39.6	50.6	65.2	71.0	73.3	60.0
	ReSET	43.8	54.0	64.9	72.1	77.5	**62.5**
GPQA-Diamond	BF16 Baseline	48.3	58.4	54.4	61.3	64.1	57.3
	RTN	47.1	53.7	50.7	57.4	62.8	54.3
	BRQ	45.5	56.7	53.5	58.2	60.8	55.0
	4/6	44.6	55.4	51.9	57.6	64.6	54.8
	MR-GPTQ	43.4	57.4	52.4	60.9	60.6	54.9
	ReSET	46.0	57.6	53.2	58.5	62.9	**55.6**
LiveCodeBench	BF16 Baseline	28.2	38.5	43.2	46.5	49.9	41.3
	RTN	29.5	37.1	36.4	46.7	46.5	39.2
	BRQ	27.6	37.7	36.2	39.0	45.1	37.1
	4/6	27.0	36.4	35.8	45.3	45.7	38.0
	MR-GPTQ	27.2	34.9	42.9	47.2	45.1	39.5
	ReSET	28.4	37.9	42.1	46.1	46.7	**40.2**

Best mean accuracy among the evaluated NVFP4 methods

Gains over RTN: +2.6 pp (AIME), +1.3 pp (GPQA), +1.0 pp (LiveCodeBench)

NVFP4 Kernels for Small-Batch Decoding

CUDA-core execution matches the actual latency-critical scenario



128-row tile in the evaluated Tensor-Core path

Register-only FP16 dequantization with half2 FMA
Tensor Cores handle prefill and larger decode batches

Kernel-Level Latency Evaluation

Projection kernels latency (µs)
Qwen3-8B, 14B and 32B

Qwen3	N	K	vLLM	**Ours**	Speedup
8B	6144	4096	8.69	4.28	2.03×
	4096	4096	7.69	3.75	2.05×
	12288	4096	11.45	6.40	1.79×
	4096	12288	16.61	6.66	2.49×
14B	7168	5120	10.79	5.53	1.95×
	5120	5120	9.08	4.64	1.96×
	17408	5120	14.93	9.49	1.57×
	5120	17408	23.77	10.35	2.30×
32B	10240	5120	13.41	6.62	2.03×
	5120	8192	12.40	7.73	1.60×
	25600	5120	25.89	13.25	1.96×
	5120	25600	33.26	15.00	2.22×

Pipeline breakdown (µs)
Qwen3-32B

Layer Type	vLLM		MR-GPTQ		**Ours**	
	Q	GEMM	Q+R	GEMM	Q	GEMV
QKV	1.81	13.41	4.65	11.84	**1.65**	**6.62**
Out	2.04	12.40	4.87	11.90	**1.81**	**7.73**
Gate-Up	1.83	25.89	4.55	25.90	**1.67**	**13.25**
Down	2.80	33.26	5.08	33.87	**1.79**	**15.00**

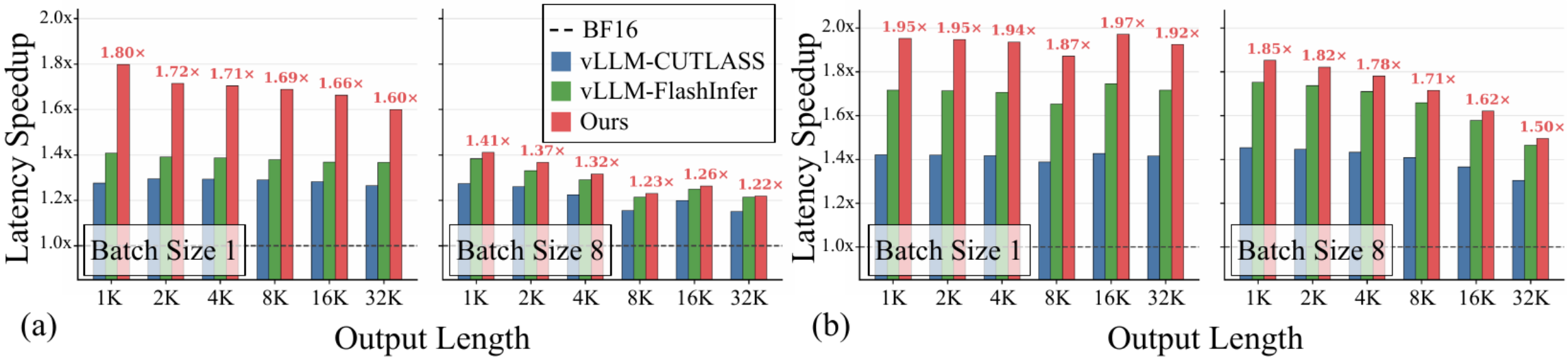
No additional rotation as in MR-GPTQ

1.57–2.49× faster projection kernels

B200, M = 1. Projection-kernel speedup vs. vLLM-CUTLASS NVFP4.

End-to-End Decoding Speedup

Full-system speedup over BF16, including the ReSET sampler



(a) Qwen3-8B

(b) Qwen3-32B

Up to 1.97×

~1.5% sampling overhead

~100 μs/token, included in the measurement

Single B200, input 512 tokens, batch sizes 1 and 8, output lengths 1K–32K

Conclusion

Step-aware temperature scaling

+2.6 percentage points on AIME-120 over NVFP4 RTN
No model retraining

Small-batch NVFP4 kernels

Up to 1.97× end-to-end decoding speedup over BF16

Thank You !



Code and implementation

github.com/aiha-lab/ReSET