A SUPPLEMENTARY EXPERIMENTS

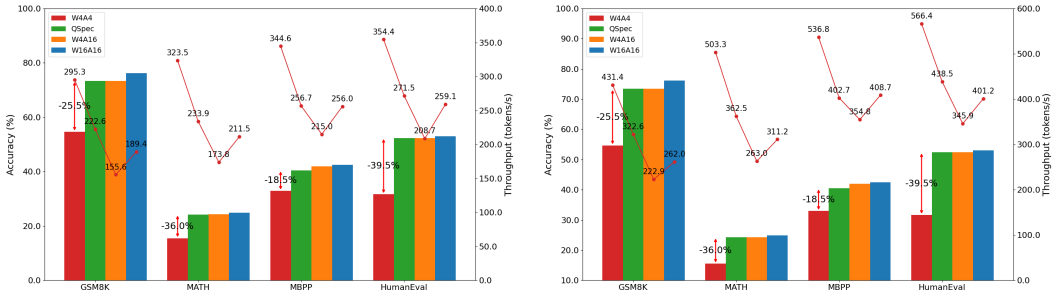
We further extend our experiments on NVIDIA L20 GPUs, and complement additional analysis of W16A16 (Wolf et al., 2020), Atom-based W4A16 (Lin et al., 2024a), W4A4 (Zhao et al., 2024b), and QSPEC.

Consistent Efficiency Enhancement of QSPEC over W4A16. As presented in Table 5, we detail the token generation throughput for both QSPEC and WXAX methods across various model sizes, quantization configurations, batch sizes, and datasets. Compared to W4A16, QSPEC achieves a throughput increase of $1.33\times$ across all the settings on average, with a peak improvement of $1.64\times$. These results, along with those in Table 4, validate the consistent efficiency superiority of QSPEC over W4A16 on different GPU platforms. Additionally, QSPEC consistently outperforms W16A16 in terms of efficiency across all the settings.

Preserved Generation Quality of QSPEC Compared to W4A16. As illustrated in Figure 6, we visualize the generation quality (i.e., accuracy) and efficiency (i.e., throughput). Aligning with the analysis of Table 1, W4A4 experiences a significant performance decline, ranging from 18.5% to 39.5%, on multi-step reasoning benchmarks when compared to W4A16. In contrast, QSPEC not only maintains the performance of W4A16 (slightly lower than that of W16A16 due to weight quantization for memory saving), but also offers much higher throughput.

Detailed Latency Decomposition of Per Valid Token. As shown in Figure 7, we calculate the per-valid-token latency by dividing the total latency by the number of accepted tokens in each sample, which is then averaged across all samples and evaluation datasets. Notably, the decode stage accounts for the majority of the time latency when compared to the prefill stage. With the rapid drafting capability and parallel verification, QSPEC achieves significantly lower latency than W4A16, ranging from 28.5% to 39.7%. In detail, QSPEC spends more time in the draft phase than in the high-precision verify phase. This may be attributed to the high acceptance rate of QSPEC, which resulted in less verify requests.

Ablation on Draft Token Length. To assess parameter sensitivity, we vary the draft token length γ, the sole hyperparameter of QSPEC, from 2 to 7 across all benchmarks with Llama3.2-3b and Llama3-8b-instruct models. For a thorough comparison, we also include the throughput of W16A16 and W4A16 as references. As depicted in Figure 8, an increase in γ results in a gradual decrease in the token acceptance rate, since the rejection of any token leads to the discarding of all subsequent tokens. Nevertheless, even at $\gamma = 7$, the token acceptance rate remains relatively high at approximately 70%, compared to the 28%–58% observed in the 160m–7b draft-target model pair under $\gamma = 5$ in conventional speculative decoding (Liu et al., 2024). Additionally, we observe a continuous improvement in throughput compared to W4A16, indicating the hyperparameter robustness of QSPEC. With an appropriate choice of γ (*i.e.*, $\gamma \leq 5$), QSPEC consistently outperforms W16A16 in both memory consumption and efficiency.



(a) Batch Size 8

(b) Batch Size 16

Figure 6: Comparison of accuracy and efficiency among W16A16, W4A16, W4A4, and QSPEC across various datasets with batch sizes of 8 and 16, respectively. The bars and lines represent the accuracy and throughput of each method.

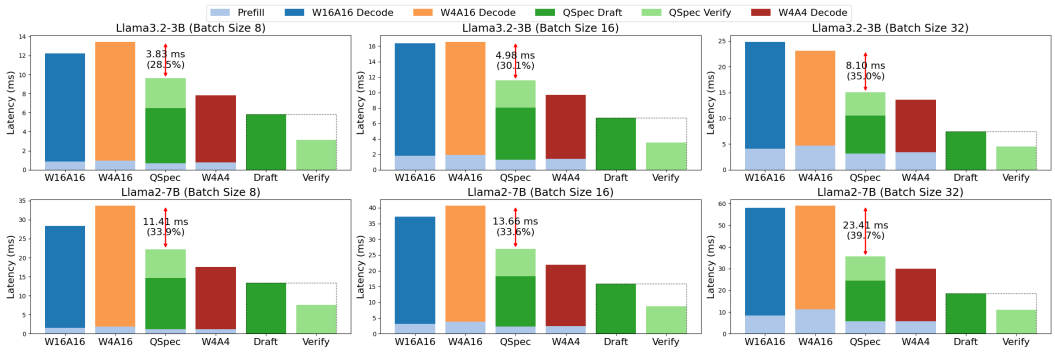
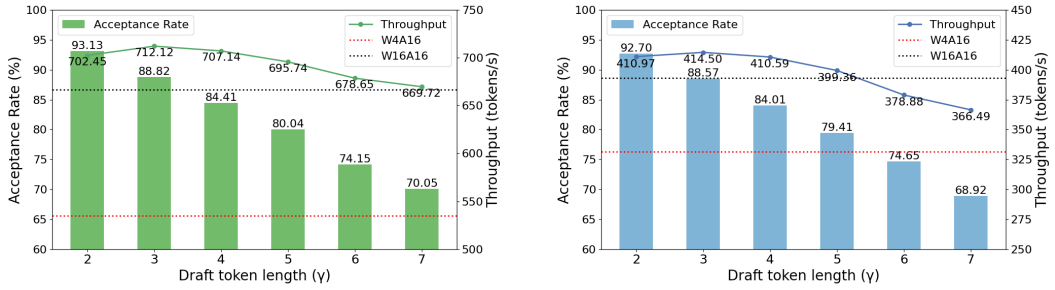


Figure 7: Per-valid-token latency decomposition of W16A16, W4A16, QSPEC and W4A4 across different models and batch sizes. The latency of QSPEC is further decomposed into draft and verify categories for details.



(a) Llama3.2-3b

(b) Llama3-8b-instruct

Figure 8: Acceptance rate and throughput of Llama 3.2-3b (with a batch size of 8) and Llama 3-8b-instruct (with a batch size of 16) with respect to the draft token length γ.

Table 5: Comparison of token generation throughput across different model sizes, quantization configurations, and batch sizes for various datasets. All values are measured in token/s. "Avg." denotes the average speedup ratio for the corresponding row or column. "†" indicates the failure of W4A16 kernels to support these batch sizes together with long sequences and the large models.

Model	Method	Batch	GSM8K	MATH	MBPP	HumanEval	ShareGPT	LMsys-1k	Avg.
3B[1]	W16A16	8	511.1	588.7	756.6	647.2	785.7	711.2	–
		16	666.5	845.6	1171.0	948.3	1292.2	1126.4	–
		32	833.4	1081.5	1697.7	1111.6	1975.6	1553.3	–
	W4A4	8	804.7	921.2	1002.0	892.6	1091.6	990.3	–
		16	1109.1	1374.5	1548.0	1289.8	1763.5	1581.0	–
		32	1424.3	1899.3	2300.6	1488.2	2777.3	2194.4	–
	W4A16	8	420.0	476.7	604.5	535.7	610.4	559.8	–
		16	578.5	715.9	989.7	804.4	1080.2	925.8	–
		32	726.3	933.8	1536.7	954.4	1704.5	1336.4	–
	QSpec	8	594.1 (1.41×)	648.2 (1.36×)	760.1 (1.26×)	723.6 (1.35×)	787.5 (1.29×)	738.8 (1.32×)	1.33×
		16	811.5 (1.40×)	936.0 (1.31×)	1157.8 (1.17×)	1042.1 (1.30×)	1294.5 (1.20×)	1171.4 (1.27×)	1.27×
		32	1030.4 (1.42×)	1240.2 (1.33×)	1617.4 (1.05×)	1248.5 (1.31×)	1969.6 (1.16×)	1576.0 (1.18×)	1.24×
		Avg.	1.41×	1.33×	1.16×	1.32×	1.21×	1.25 ×	1.28×
7B	W16A16	8	213.4	254.3	278.8	316.7	322.4	285.3	–
		16	290.3	362.1	447.7	505.1	541.3	441.6	–
		32	340.9	441.6	585.3	663.6	735.3	564.2	–
	W4A4	8	349.5	411.7	396.1	471.2	471.8	419.4	–
		16	496.6	612.2	614.3	749.5	760.9	642.6	–
		32	620.0	793.6	801.5	1043.9	1083.2	865.5	–
	W4A16	8	165.0	193.1	224.5	240.2	243.5	220.2	–
		16	231.8	286.5	384.4	407.3	435.9	358.0	–
		32	268.9	359.9	480.0	555.9	620.2	470.1	–
	QSpec	8	253.7 (1.54×)	291.5 (1.51×)	298.3 (1.33×)	350.9 (1.46×)	345.7 (1.42×)	310.3 (1.41×)	1.44×
		16	359.8 (1.55×)	420.2 (1.47×)	466.7 (1.21×)	555.2 (1.36×)	557.8 (1.28×)	473.1 (1.32×)	1.37×
		32	441.8 (1.64×)	527.2 (1.46×)	575.3 (1.20×)	749.4 (1.35×)	770.0 (1.24×)	628.4 (1.34×)	1.39×
		Avg.	1.58×	1.48×	1.25×	1.39×	1.31×	1.36×	1.39×
8B	W16A16	8	189.4	211.5	256.0	259.1	290.7	265.8	–
		16	262.0	311.2	408.7	401.2	511.0	447.4	–
		32	303.8	390.8	566.3	522.6	820.0	649.8	–
	W4A4	8	295.3	323.5	344.6	354.4	395.9	366.8	–
		16	431.4	503.3	536.8	566.4	697.5	621.1	–
		32	532.8	688.5	755.7	763.7	1167.9	956.8	–
	W4A16	8	155.6	173.8	215.0	208.7	231.1	215.6	–
		16	222.9	263.0	354.8	345.9	422.8	369.4	–
		32	†	†	509.8	468.7	706.0	580.5	–
	QSpec	8	222.6 (1.43×)	233.9 (1.35×)	256.7 (1.19×)	271.5 (1.30×)	285.0 (1.23×)	268.3 (1.24×)	1.29×
		16	322.6 (1.45×)	362.5 (1.38×)	402.7 (1.14×)	438.5 (1.27×)	507.5 (1.20×)	453.5 (1.23×)	1.28×
		32	400.2 (†)	362.5 (†)	578.1 (1.13×)	573.0 (1.22×)	798.8 (1.13×)	684.5 (1.18×)	1.27×
		Avg.	1.44×	1.36×	1.15×	1.26×	1.19×	1.22×	1.27 ×
13B[1]	W16A16	8	121.9	146.6	183.1	182.0	187.1	160.1	–
		16	169.6	211.2	304.4	291.0	311.0	243.0	–
		32	202.4	253.8	426.0	423.5	311.0	334.2	–
	W4A4	8	194.7	228.2	253.6	261.5	259.8	228.2	–
		16	288.3	349.2	415.3	424.9	431.5	348.4	–
		32	369.8	469.9	606.7	665.4	431.5	508.8	–
	W4A16	8	94.8	112.9	143.4	140.0	146.7	127.9	–
		16	136.1	171.9	250.8	236.9	255.9	207.2	–
		32	†	†	376.4	365.5	255.9	287.4	–
	QSpec	8	148.2 (1.56×)	167.9 (1.49×)	193.6 (1.35×)	201.2 (1.44×)	194.5 (1.33×)	174.0 (1.36×)	1.42×
		16	212.8 (1.56×)	248.6 (1.45×)	316.8 (1.26×)	323.3 (1.36×)	327.4 (1.28×)	266.9 (1.29×)	1.29×
		32	266.6 (†)	320.0 (†)	451.5 (1.20×)	483.0 (1.32×)	327.4 (1.28×)	379.3 (1.32×)	1.32×
		Avg.	1.56×	1.47×	1.27×	1.37×	1.29×	1.32×	1.38×