

# DGX Spark GB10 模型推理性能基准报告

Blade-LLM · Illusionna

14:00, Tuesday 30<sup>th</sup> June, 2026

## 目录

<b>1</b>	<b>参数解释</b>	<b>1</b>
<b>2</b>	<b>总结报告</b>	<b>1</b>
2.1	超长历史上文吞吐延迟测试	1
2.2	一般指标预览	1
2.3	对比总结	3
<b>3</b>	<b>模型测评</b>	<b>3</b>
3.1	DeepSeek-V4-Flash	3
3.1.1	测试环境 ds4 + DeepSeek-V4-Flash-IQ2XXS	3
3.1.2	测试环境 ds4 + DeepSeek-V4-Flash-IQ2XXS + MTP-2	4
3.1.3	生产环境 ds4 + DeepSeek-V4-Flash-IQ2XXS + MTP-2	5
3.2	Qwen3.6-27B	6
3.2.1	SGLang.build + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-2	6
3.2.2	SGLang.build + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-4	8
3.2.3	SGLang.build + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-8	9
3.2.4	SGLang.pull + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-2	11
3.2.5	SGLang.pull + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-4	12
3.2.6	SGLang.pull + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-8	14
3.2.7	SGLang.runtime + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-2	15
3.2.8	SGLang.runtime + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-4	17
3.2.9	SGLang.runtime + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-8	18
3.2.10	vLLM.dev + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-2	20
3.2.11	vLLM.dev + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-4	21
3.2.12	vLLM.dev + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-8	23
3.2.13	vLLM.build + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-2	24
3.2.14	vLLM.build + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-4	26
3.2.15	vLLM.build + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-8	27
3.2.16	vLLM.build + Qwen3.6-27B-INT4-AutoRound-Intel + DFlash-2	29
3.2.17	vLLM.build + Qwen3.6-27B-INT4-AutoRound-Intel + DFlash-4	30
3.2.18	vLLM.build + Qwen3.6-27B-INT4-AutoRound-Intel + DFlash-8	32
3.2.19	vLLM.build + Qwen3.6-27B-INT4-AutoRound-Intel + DFlash-15	33

3.2.20	vLLM.dflash + Qwen3.6-27B-INT4-AutoRound-Intel + DFlash-15 . . . . .	35
3.2.21	LLaMA.cpp + Qwen3.6-27B-Q4_K_M + MTP-2 . . . . .	36
3.2.22	LLaMA.cpp + Qwen3.6-27B-Q4_K_M + MTP-4 . . . . .	37
3.2.23	LLaMA.cpp + Qwen3.6-27B-Q4_K_M + MTP-8 . . . . .	39
3.3	Qwen3.5-122B . . . . .	40
3.3.1	vLLM.dev + Qwen3.5-122B-A10B-INT4-AutoRound-Intel + MTP-2 . . . . .	40
3.3.2	vLLM.dev + Qwen3.5-122B-A10B-INT4-AutoRound-Intel + MTP-4 . . . . .	41
3.3.3	vLLM.dev + Qwen3.5-122B-A10B-INT4-AutoRound-Intel + MTP-8 . . . . .	43
3.3.4	vLLM.dflash + Qwen3.5-122B-A10B-INT4-AutoRound-Intel + DFlash-2 . . . . .	44
3.3.5	vLLM.dflash + Qwen3.5-122B-A10B-INT4-AutoRound-Intel + DFlash-8 . . . . .	46
3.3.6	vLLM.dflash + Qwen3.5-122B-A10B-INT4-AutoRound-Intel + DFlash-15 . . . . .	47
3.3.7	vLLM.dev + Qwen3.5-122B-Hybrid-INT4FP8 + MTP-2 . . . . .	49
3.3.8	vLLM.dev + Qwen3.5-122B-Hybrid-INT4FP8 + MTP-4 . . . . .	50
3.3.9	vLLM.dev + Qwen3.5-122B-Hybrid-INT4FP8 + MTP-8 . . . . .	52
3.4	补充测试 . . . . .	53
3.4.1	vLLM.dev + Qwen3.6-27B . . . . .	53
3.4.2	vLLM.dflash + Qwen3.6-27B-NVFP4 + DFlash-8 . . . . .	55
3.4.3	vLLM.dflash + Qwen3.6-27B-NVFP4 + DFlash-15 . . . . .	56
3.4.4	vLLM.dflash + Qwen3.6-27B-AWQ-INT4 + DFlash-15 . . . . .	58
3.4.5	vLLM.dflash + Qwen3.6-27B-GPTQ-INT4 + DFlash-15 . . . . .	59
3.4.6	vLLM.dflash + Qwen3.6-27B-FP8 + DFlash-15 . . . . .	61
3.4.7	vLLM.dflash + Qwen3.6-27B-AEON-NVFP4-MTP-XS + DFlash-15 . . . . .	62
3.4.8	LLaMA.cpp + Qwen3.6-27B-UD-Q4_K_XL + MTP-2 . . . . .	64
3.4.9	Qwen3.6-27B-ParoQuant + NVFP4 . . . . .	65
3.4.10	Qwen3.6-27B-TQ3_4S + MTP . . . . .	65
3.4.11	vLLM.dev + Qwen3.5-122B-A10B-INT4-AutoRound-Intel + MTP-2 + Long-Text . . . . .	65
3.4.12	vLLM.dev + Qwen3.5-122B-Hybrid-INT4FP8 + MTP-2 + Long-Text . . . . .	65



# 1 参数解释

- **测试服务器**: ssh ai@192.168.120.32
- **pp**: 填充 token 数量
- **tg**: 生成 token 数量
- **c**: 并发请求数
- **d**: 已缓存历史上文 token 数量
- **pp t/s**: 提示词（预填充）阶段，每秒处理多少 token
- **tg t/s**: 解码生成阶段，每秒生成多少 token
- **TTFT (ms)**: 从发出请求到收到第一个 token 的延迟
- **Total (ms)**: 从请求发出，到完整响应全部生成完毕的总时间
- **Warm-up (ms)**: 预热所需时间
- **Median-turn (s)**: 所有对话轮次耗时的中位数
- **Quality**: 生成内容质量
- **Responsiveness**: 响应质量性能

# 2 总结报告

历史上文深度（d0 / d4096 / d8192）与并发数（c1 / c2 / c4），超长上文历史深度为（d131072），背景透明无色为该工况下的解码器生成速度吞吐的最优值。

## 2.1 超长历史上文吞吐延迟测试

vLLM:EngineCore		pp t/s			tg t/s			TTFT (s)			
d131072		GiB	c1	c2	c4	c1	c2	c4	c1	c2	c4
	vLLM.dev + Qwen3.5-122B-A10B-INT4-AutoRound-Intel + MTP-2	106.49	1117	1096	1085	32.4	25.2	23.1	106.8	163.5	274.6
	vLLM.dev + Qwen3.5-122B-Hybrid-INT4FP8 + MTP-2	108.93	1119	1101	1091	33.4	13.9	10.7	106.8	162.8	273.6

## 2.2 一般指标预览

Params	GiB Quality Responsiveness Warm-up Median-turn					pp t/s												tg t/s												TTFT (ms)																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																			
						d0												d4096												d8192																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																			
						c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4	c1 c2 c4

## 2.3 对比总结

- **综合最优**: ★ vLLM.dflash + Q3.5-122B-A10B-INT4 + DFlash-8, 质量满分 100, 单流 TTFT 全场最低 1.13s (d0/c1), 解码 tg 全场最高 91.7 t/s (d0/c4), 预填充 pp 亦达 2279 t/s, 中位轮次仅 3.3s, 在 TTFT、TPS、Quality 中是延迟、吞吐场景的首选。
- **响应最佳**: vLLM.dev + Q3.5-122B-A10B Hybrid-INT4FP8 + MTP-2, 质量 100、响应 48, TTFT 1.25s、pp 峰值 2421 t/s、tg 80.5 t/s、中位轮次 3.1s, 各项指标几乎贴平 vLLM.dflash + Q3.5-122B-A10B-INT4 + DFlash-8, 响应与预填充更强、峰值 tg 略低, 是更稳健的次优选择。
- **首字延迟 TTFT**: 最低的 TTFT 是 122B-A10B MoE, d0/c1 落在 1.13–1.28s, 2200–2421 t/s 的预填充吞吐, 约为 27B 稠密档 (400–1000 t/s) 的 2–5 倍, 27B 中仅 NVFP4 的 pp 能上探 1900–2000, TTFT 最低也要 1.34s。
- **解码吞吐 TPS**: 122B-A10B 的 tg 在 c4 并发普遍 75–92 t/s 且深上文衰减平缓 (d4096/c4 80、d8192/c4 52 t/s); 27B 稠密模型中仅 vLLM.build + MTP-4 (d0/c4 91.3)、NVFP4 + DFlash-8 (87.5) 能短暂逼近, 但深上文高并发下回落更快。
- **低延迟**: vLLM.dflash + Q3.6-27B-NVFP4 + DFlash-8 ——TTFT 低至 1.34s、pp 峰值 1991 t/s, 但 tg 在 30–88 间剧烈波动、响应仅 37, 深上文高并发衰减明显, 适合高 QPS 短输出。
- **投机解码**: 草稿步数并非越大越好, 122B 上 MTP-2 / DFlash-2 (响应 48) 最佳, 加大到 MTP-8 / DFlash-15 后 tg 与响应回落 (响应降至 40) 且质量骤降 (MTP-8 质量 73); 27B 上 vLLM.dev MTP-8 质量跌至 53、vLLM.build DFlash-2/4/8 质量仅 43, DFlash-8 是 122B 兼顾 TTFT、TPS 与质量的较优解, 过深步数引发质量反噬。
- **DeepSeek-V4-Flash 的 MTP 反噬**: DeepSeek-V4-Flash-IQ2XXS 单跑尚可 (pp 1100、tg 峰值 81.9、质量 90), 叠加 MTP-2 后预填充崩塌至 133–365 t/s、tg 跌至 15–21 t/s、质量降至 83、TTFT 同步抬高 (d0/c1 5.85s), 在 DGX Spark GB10 硬件上为显著负优化。
- **FP8 量化模型**: vLLM.dflash + Q3.6-27B-FP8 + DFlash-15 虽质量 100, 却随并发与深度急剧劣化——pp 由 d0/c1 的 1746 t/s 跌到 d8192/c4 的 342 t/s, TTFT 同步爆炸至 119s (d8192/c4), 稳定性最差, 不宜高并发长上文。
- **超长 128K 上文**: vLLM.dev 122B-A10B MTP-2 的 tg (c1/c2/c4 = 32.4 / 25.2 / 23.1) 在多并发下保持力远优于 Hybrid-INT4FP8 (33.4 / 13.9 / 10.7), 两者 pp 1100、TTFT 仅 107–275ms, 极长上文且多并发时 A10B 更稳。
- **小体积**: LLaMA.cpp + Q3.6-27B (Q4\_K\_M / UD-Q4\_K\_XL) 显存占用仅 38–41GiB (约为其余方案的 40%), MTP-8 档质量满分 100, 但 tg 仅 18–28 t/s、TTFT 较高 (d0/c1 2.8s) 且深上文 / 并发下进一步走低, 适合显存受限、吞吐要求不高的部署。

## 3 模型测评

### 3.1 DeepSeek-V4-Flash

#### 3.1.1 测试环境 ds4 + DeepSeek-V4-Flash-IQ2XXS

```
./ds4/ds4-server \
```

```

-m ./models/DeepSeek-V4-Flash-IQ2XXS-w2Q2K-AProjQ8-SExpQ8-OutQ8-chat-v2-imatrix.gguf \
--host 0.0.0.0 \
--port 4321 \
--ctx 32768 \
--tokens 131072 \
--kv-disk-dir ./tmp-ds4-kv \
--kv-disk-space-mb 8192 \
--warm-weights

```

```

tool-eval-bench --base-url http://127.0.0.1:4321 --short --perf

```

GPU 110.63GiB • Warm-up 8865ms • Quality 90/100 • Responsiveness 12/100 • Median turn 11.3s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
■	d0	c1	1,157	24.5	2,100	7,002
		c2	1,075	49.7	3,715	8,067
		c4	1,119	81.9	7,261	12,274
■	d4096	c1	1,167	28.1	5,588	9,821
		c2	1,122	50.4	10,826	15,289
		c4	1,130	34.6	18,860	25,110
■	d8192	c1	1,108	25.1	9,583	14,373
		c2	1,066	47.6	19,124	23,864
		c4	1,062	23.7	31,650	40,714
Per-stream (single request, per concurrent client)						
■	-	c1	411	27.5	16,343	-
		c2	296	22.1	-	-
		c4	274	19.5	-	-

### 3.1.2 测试环境 ds4 + DeepSeek-V4-Flash-IQ2XXS + MTP-2

```

./ds4/ds4-server \
-m ./models/DeepSeek-V4-Flash-IQ2XXS-w2Q2K-AProjQ8-SExpQ8-OutQ8-chat-v2-imatrix.gguf \
--mtp ./models/DeepSeek-V4-Flash-MTP-Q4K-Q8_0-F32.gguf \
--mtp-draft 2 \
--host 0.0.0.0 \
--port 4321 \
--ctx 32768 \
--tokens 65536 \
--kv-disk-dir ./tmp-ds4-kv \
--kv-disk-space-mb 16384 \
--warm-weights

```

```
tool-eval-bench --base-url http://127.0.0.1:4321 --short --perf
```

GPU 108.19GiB • Warm-up 96923ms • Quality 83/100 • Responsiveness 13/100 • Median turn 10.8s						
Depth		Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
🔵	d0	c1	365	21.1	5,850	11,241
		c2	163	17.5	14,456	19,668
		c4	133	15.8	31,574	37,057
🟡	d4096	c1	355	20.7	16,330	21,996
		c2	253	13.7	30,766	35,740
		c4	224	10.1	59,151	64,890
🟢	d8192	c1	336	19.8	28,715	34,581
		c2	275	9.3	47,925	53,665
		c4	254	7.7	88,481	94,061
Per-stream (single request, per concurrent client)						
🟠	—	c1	365	21.1	5,850	—
		c2	275	17.5	—	—
		c4	254	15.8	—	—

### 3.1.3 生产环境 ds4 + DeepSeek-V4-Flash-IQ2XXS + MTP-2

```

IMAGE="i-dgx-spark-gb10:ds4flash"

MODEL_PATH="./models/DeepSeek-V4-Flash-IQ2XXS-w2Q2K-AProjQ8-SExpQ8-OutQ8-chat-v2-imatrix.gguf"
MTP_PATH="./models/DeepSeek-V4-Flash-MTP-Q4K-Q8_0-F32.gguf"

MODEL_NAME=$(basename "$MODEL_PATH")
MTP_NAME=$(basename "$MTP_PATH")

docker run \
  --rm -it \
  --gpus all \
  --name "DeepSeek-V4-Flash" \
  -v "${MODEL_PATH}:/workspace/${MODEL_NAME}" \
  -v "${MTP_PATH}:/workspace/${MTP_NAME}" \
  -p 4321:4321 \
  --ipc=host \
  --entrypoint /workspace/ds4/ds4-server \
  ${IMAGE} \
  -m "/workspace/${MODEL_NAME}" \

```

```
--mtp "/workspace/${MTP_NAME}" \
--mtp-draft 2 \
--host 0.0.0.0 \
--port 4321 \
--ctx 32768 \
--tokens 65536 \
--kv-disk-dir /tmp/ds4-kv \
--kv-disk-space-mb 16384 \
--warm-weights
```

```
tool-eval-bench --base-url http://127.0.0.1:4321 --short --perf
```

GPU 108.17GiB • Warm-up 96522ms • Quality 90/100 • Responsiveness 12/100 • Median turn 11.0s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
<i>Aggregate throughput (pp2048 / tg128)</i>						
■	d0	c1	380	20.8	5,798	11,260
		c2	171	17.1	13,659	19,308
		c4	134	15.8	30,839	36,411
■	d4096	c1	346	21.0	17,070	22,580
		c2	249	14.2	31,189	36,064
		c4	223	10.2	59,386	65,045
■	d8192	c1	342	20.9	28,357	33,950
		c2	275	9.9	48,611	54,118
		c4	255	8.0	88,756	94,069
<i>Per-stream (single request, per concurrent client)</i>						
■	—	c1	380	21.0	17,070	—
		c2	275	17.1	—	—
		c4	255	15.8	—	—

## 3.2 Qwen3.6-27B

### 3.2.1 SGLang.build + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-2

```
MODEL_PATH="./models/Qwen3.6-27B-INT4-AutoRound-Intel"
MODEL_NAME=$(basename "$MODEL_PATH")
SGLANG_IMAGE="i-dgx-spark-gb10:sglang.build"

docker run \
  --rm -it \
  --gpus all \
  --name ${MODEL_NAME} \
```



```

-v ${MODEL_PATH}:/models/${MODEL_NAME} \
-p 4321:4321 \
--ipc=host \
-e SGLANG_ENABLE_SPEC_V2=1 \
--health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
--health-interval 10s \
--health-timeout 5s \
--health-retries 60 \
--health-start-period 900s \
${SGLANG_IMAGE} sglang serve \
  --model-path /models/${MODEL_NAME} \
  --served-model-name ${MODEL_NAME} \
  --host 0.0.0.0 \
  --port 4321 \
  --mem-fraction-static 0.75 \
  --context-length 262144 \
  --max-running-requests 16 \
  --cuda-graph-max-bs 16 \
  --max-prefill-tokens 16384 \
  --chunked-prefill-size 8192 \
  --quantization auto-round \
  --kv-cache-dtype fp8_e5m2 \
  --reasoning-parser qwen3 \
  --tool-call-parser qwen3_coder \
  --attention-backend flashinfer \
  --mamba-scheduler-strategy extra_buffer \
  --speculative-algorithm NEXTN \
  --speculative-num-steps 2 \
  --speculative-eagle-topk 1 \
  --speculative-num-draft-tokens 3 \
  --trust-remote-code

HF_HUB_OFFLINE=1 tool-eval-bench \
  --base-url http://127.0.0.1:4321 \
  --short --perf \
  --bench-args "--tokenizer ./models/Qwen3.6-27B-INT4-AutoRound-Intel"

```

GPU 89.65GiB • Warm-up 1446ms • Quality 97/100 • Responsiveness 22/100 • Median turn 6.9s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
❑	d0	c1	402	17.0	5,566	12,668
		c2	402	29.8	10,005	17,673
		c4	414	56.6	19,669	27,693
❑	d4096	c1	416	16.5	15,183	22,500
		c2	412	29.7	29,588	37,240
		c4	410	14.8	52,135	63,760
❑	d8192	c1	405	16.8	25,739	32,948
		c2	403	30.3	50,660	58,294
		c4	406	10.1	82,728	99,411
Per-stream (single request, per concurrent client)						
❑	—	c1	416	17.0	5,566	—
		c2	412	30.3	—	—
		c4	414	56.6	—	—

### 3.2.2 SGLang.build + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-4

```

MODEL_PATH="./models/Qwen3.6-27B-INT4-AutoRound-Intel"
MODEL_NAME=$(basename "$MODEL_PATH")
SGLANG_IMAGE="i-dgx-spark-gb10:sglang.build"

docker run \
  --rm -it \
  --gpus all \
  --name ${MODEL_NAME} \
  -v ${MODEL_PATH}:/models/${MODEL_NAME} \
  -p 4321:4321 \
  --ipc=host \
  -e SGLANG_ENABLE_SPEC_V2=1 \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  ${SGLANG_IMAGE} sglang serve \
    --model-path /models/${MODEL_NAME} \
    --served-model-name ${MODEL_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --mem-fraction-static 0.75 \

```

```

--context-length 262144 \
--max-running-requests 16 \
--cuda-graph-max-bs 16 \
--max-prefill-tokens 16384 \
--chunked-prefill-size 8192 \
--quantization auto-round \
--kv-cache-dtype fp8_e5m2 \
--reasoning-parser qwen3 \
--tool-call-parser qwen3_coder \
--attention-backend flashinfer \
--mamba-scheduler-strategy extra_buffer \
--speculative-algorithm NEXTN \
--speculative-num-steps 4 \
--speculative-eagle-topk 1 \
--speculative-num-draft-tokens 5 \
--trust-remote-code

```

```

HF_HUB_OFFLINE=1 tool-eval-bench \
--base-url http://127.0.0.1:4321 \
--short --perf \
--bench-args "--tokenizer ./models/Qwen3.6-27B-INT4-AutoRound-Intel"

```

GPU 90.00GiB • Warm-up 12474ms • Quality 97/100 • Responsiveness 26/100 • Median turn 6.0s						
Depth		Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
❑	d0	c1	412	20.8	5,483	11,166
		c2	403	29.5	10,014	17,018
		c4	412	53.0	19,765	27,444
❑	d4096	c1	416	20.9	15,240	20,904
		c2	411	30.8	29,647	36,456
		c4	410	14.6	52,037	63,517
❑	d8192	c1	405	18.8	25,736	32,061
		c2	401	33.0	50,935	57,632
		c4	404	10.1	83,137	99,022
Per-stream (single request, per concurrent client)						
❑	—	c1	416	20.9	15,240	—
		c2	411	33.0	—	—
		c4	412	53.0	—	—

### 3.2.3 SGLang.build + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-8

```

MODEL_PATH="./models/Qwen3.6-27B-INT4-AutoRound-Intel"
MODEL_NAME=$(basename "$MODEL_PATH")
SGLANG_IMAGE="i-dgx-spark-gb10:sglang.build"

docker run \
  --rm -it \
  --gpus all \
  --name ${MODEL_NAME} \
  -v ${MODEL_PATH}:/models/${MODEL_NAME} \
  -p 4321:4321 \
  --ipc=host \
  -e SGLANG_ENABLE_SPEC_V2=1 \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  ${SGLANG_IMAGE} sglang serve \
    --model-path /models/${MODEL_NAME} \
    --served-model-name ${MODEL_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --mem-fraction-static 0.75 \
    --context-length 262144 \
    --max-running-requests 16 \
    --cuda-graph-max-bs 16 \
    --max-prefill-tokens 16384 \
    --chunked-prefill-size 8192 \
    --quantization auto-round \
    --kv-cache-dtype fp8_e5m2 \
    --reasoning-parser qwen3 \
    --tool-call-parser qwen3_coder \
    --attention-backend flashinfer \
    --mamba-scheduler-strategy extra_buffer \
    --speculative-algorithm NEXTN \
    --speculative-num-steps 8 \
    --speculative-eagle-topk 1 \
    --speculative-num-draft-tokens 9 \
    --trust-remote-code

HF_HUB_OFFLINE=1 tool-eval-bench \
  --base-url http://127.0.0.1:4321 \
  --short --perf \
  --bench-args "--tokenizer ./models/Qwen3.6-27B-INT4-AutoRound-Intel"

```

GPU 90.32GiB • Warm-up 10201ms • Quality 97/100 • Responsiveness 27/100 • Median turn 5.9s						
	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
❑	d0	c1	394	16.1	5,792	13,182
		c2	401	28.2	10,020	17,520
		c4	403	42.4	20,179	29,206
❑	d4096	c1	417	17.7	15,289	21,968
		c2	410	35.7	29,731	35,790
		c4	394	13.8	54,343	67,017
❑	d8192	c1	406	17.6	25,798	32,516
		c2	390	30.0	52,256	59,507
		c4	405	9.7	82,832	101,635
Per-stream (single request, per concurrent client)						
❑	—	c1	417	17.7	15,289	—
		c2	410	35.7	—	—
		c4	405	42.4	—	—

### 3.2.4 SGLang.pull + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-2

```

MODEL_PATH="./models/Qwen3.6-27B-INT4-AutoRound-Intel"
MODEL_NAME=$(basename "$MODEL_PATH")
SGLANG_IMAGE="i-dgx-spark-gb10:sglang.pull"

docker run \
  --rm -it \
  --gpus all \
  --name ${MODEL_NAME} \
  -v ${MODEL_PATH}:/models/${MODEL_NAME} \
  -p 4321:4321 \
  --ipc=host \
  -e SGLANG_ENABLE_SPEC_V2=1 \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  ${SGLANG_IMAGE} sglang serve \
    --model-path /models/${MODEL_NAME} \
    --served-model-name ${MODEL_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --mem-fraction-static 0.75 \

```

```

--context-length 262144 \
--max-running-requests 16 \
--cuda-graph-max-bs 16 \
--max-prefill-tokens 16384 \
--chunked-prefill-size 8192 \
--quantization auto-round \
--kv-cache-dtype fp8_e5m2 \
--reasoning-parser qwen3 \
--tool-call-parser qwen3_coder \
--attention-backend flashinfer \
--mamba-scheduler-strategy extra_buffer \
--speculative-algorithm NEXTN \
--speculative-num-steps 2 \
--speculative-eagle-topk 1 \
--speculative-num-draft-tokens 3 \
--trust-remote-code

```

```

HF_HUB_OFFLINE=1 tool-eval-bench \
--base-url http://127.0.0.1:4321 \
--short --perf \
--bench-args "--tokenizer ./models/Qwen3.6-27B-INT4-AutoRound-Intel"

```

GPU 89.01GiB • Warm-up 1349ms • Quality 97/100 • Responsiveness 23/100 • Median turn 6.8s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
<i>Aggregate throughput (pp2048 / tg128)</i>						
❑	d0	c1	415	17.7	<b>5,380</b>	<b>12,197</b>
		c2	416	27.9	9,668	17,420
		c4	<b>427</b>	<b>53.8</b>	19,092	27,301
❑	d4096	c1	424	17.1	14,911	21,995
		c2	425	31.0	28,675	35,912
		c4	423	15.3	50,439	61,646
❑	d8192	c1	416	17.3	25,020	32,022
		c2	414	31.3	49,342	56,730
		c4	415	10.3	81,006	97,721
<i>Per-stream (single request, per concurrent client)</i>						
❑	—	c1	424	17.7	5,380	—
		c2	425	31.3	—	—
		c4	427	53.8	—	—

### 3.2.5 SGLang.pull + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-4

```

MODEL_PATH="./models/Qwen3.6-27B-INT4-AutoRound-Intel"
MODEL_NAME=$(basename "$MODEL_PATH")
SGLANG_IMAGE="i-dgx-spark-gb10:sglang.pull"

docker run \
  --rm -it \
  --gpus all \
  --name ${MODEL_NAME} \
  -v ${MODEL_PATH}:/models/${MODEL_NAME} \
  -p 4321:4321 \
  --ipc=host \
  -e SGLANG_ENABLE_SPEC_V2=1 \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  ${SGLANG_IMAGE} sglang serve \
    --model-path /models/${MODEL_NAME} \
    --served-model-name ${MODEL_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --mem-fraction-static 0.75 \
    --context-length 262144 \
    --max-running-requests 16 \
    --cuda-graph-max-bs 16 \
    --max-prefill-tokens 16384 \
    --chunked-prefill-size 8192 \
    --quantization auto-round \
    --kv-cache-dtype fp8_e5m2 \
    --reasoning-parser qwen3 \
    --tool-call-parser qwen3_coder \
    --attention-backend flashinfer \
    --mamba-scheduler-strategy extra_buffer \
    --speculative-algorithm NEXTN \
    --speculative-num-steps 4 \
    --speculative-eagle-topk 1 \
    --speculative-num-draft-tokens 5 \
    --trust-remote-code

HF_HUB_OFFLINE=1 tool-eval-bench \
  --base-url http://127.0.0.1:4321 \
  --short --perf \
  --bench-args "--tokenizer ./models/Qwen3.6-27B-INT4-AutoRound-Intel"

```

GPU 90.65GiB • Warm-up 1296ms • Quality 97/100 • Responsiveness 27/100 • Median turn 5.9s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
❑	d0	c1	426	20.2	5,345	11,199
		c2	412	30.4	9,817	16,948
		c4	378	56.8	21,611	28,952
❑	d4096	c1	415	19.6	15,317	21,351
		c2	407	31.4	29,967	36,644
		c4	415	15.0	51,483	61,923
❑	d8192	c1	412	19.4	25,366	31,479
		c2	412	34.2	49,455	56,016
		c4	412	10.4	81,594	96,860
Per-stream (single request, per concurrent client)						
❑	—	c1	426	20.2	5,345	—
		c2	412	34.2	—	—
		c4	415	56.8	—	—

### 3.2.6 SGLang.pull + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-8

```

MODEL_PATH="./models/Qwen3.6-27B-INT4-AutoRound-Intel"
MODEL_NAME=$(basename "$MODEL_PATH")
SGLANG_IMAGE="i-dgx-spark-gb10:sglang.pull"

docker run \
  --rm -it \
  --gpus all \
  --name ${MODEL_NAME} \
  -v ${MODEL_PATH}:/models/${MODEL_NAME} \
  -p 4321:4321 \
  --ipc=host \
  -e SGLANG_ENABLE_SPEC_V2=1 \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  ${SGLANG_IMAGE} sglang serve \
    --model-path /models/${MODEL_NAME} \
    --served-model-name ${MODEL_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --mem-fraction-static 0.75 \

```



```

--context-length 262144 \
--max-running-requests 16 \
--cuda-graph-max-bs 16 \
--max-prefill-tokens 16384 \
--chunked-prefill-size 8192 \
--quantization auto-round \
--kv-cache-dtype fp8_e5m2 \
--reasoning-parser qwen3 \
--tool-call-parser qwen3_coder \
--attention-backend flashinfer \
--mamba-scheduler-strategy extra_buffer \
--speculative-algorithm NEXTN \
--speculative-num-steps 8 \
--speculative-eagle-topk 1 \
--speculative-num-draft-tokens 9 \
--trust-remote-code

```

```

HF_HUB_OFFLINE=1 tool-eval-bench \
--base-url http://127.0.0.1:4321 \
--short --perf \
--bench-args "--tokenizer ./models/Qwen3.6-27B-INT4-AutoRound-Intel"

```

GPU 90.56GiB • Warm-up 1574ms • Quality 97/100 • Responsiveness 26/100 • Median turn 6.0s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
<i>Aggregate throughput (pp2048 / tg128)</i>						
❑	d0	c1	422	18.5	<b>5,488</b>	<b>11,835</b>
		c2	408	25.1	9,819	17,308
		c4	422	<b>47.6</b>	19,272	27,611
❑	d4096	c1	<b>434</b>	13.3	14,708	23,747
		c2	424	34.1	28,718	35,187
		c4	426	15.0	50,129	61,411
❑	d8192	c1	425	18.8	24,634	30,882
		c2	420	30.1	48,495	55,607
		c4	422	10.1	79,365	96,546
<i>Per-stream (single request, per concurrent client)</i>						
❑	—	c1	434	18.8	24,634	—
		c2	424	34.1	—	—
		c4	426	47.6	—	—

### 3.2.7 SGLang.runtime + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-2

```

MODEL_PATH="./models/Qwen3.6-27B-INT4-AutoRound-Intel"
MODEL_NAME=$(basename "$MODEL_PATH")
SGLANG_IMAGE="i-dgx-spark-gb10:sglang.runtime"

docker run \
    --rm -it \
    --gpus all \
    --name ${MODEL_NAME} \
    -v ${MODEL_PATH}:/models/${MODEL_NAME} \
    -p 4321:4321 \
    --ipc=host \
    -e SGLANG_ENABLE_SPEC_V2=1 \
    --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
    --health-interval 10s \
    --health-timeout 5s \
    --health-retries 60 \
    --health-start-period 900s \
    ${SGLANG_IMAGE} python3 -m sglang.launch_server \
        --model-path /models/${MODEL_NAME} \
        --served-model-name ${MODEL_NAME} \
        --host 0.0.0.0 \
        --port 4321 \
        --mem-fraction-static 0.75 \
        --context-length 262144 \
        --max-running-requests 16 \
        --cuda-graph-max-bs 16 \
        --max-prefill-tokens 16384 \
        --chunked-prefill-size 8192 \
        --quantization auto-round \
        --kv-cache-dtype fp8_e5m2 \
        --reasoning-parser qwen3 \
        --tool-call-parser qwen3_coder \
        --attention-backend flashinfer \
        --mamba-scheduler-strategy extra_buffer \
        --speculative-algorithm NEXTN \
        --speculative-num-steps 2 \
        --speculative-eagle-topk 1 \
        --speculative-num-draft-tokens 3 \
        --trust-remote-code

HF_HUB_OFFLINE=1 tool-eval-bench \
    --base-url http://127.0.0.1:4321 \
    --short --perf \
    --bench-args "--tokenizer ./models/Qwen3.6-27B-INT4-AutoRound-Intel"

```

GPU 89.85GiB • Warm-up 2210ms • Quality 97/100 • Responsiveness 22/100 • Median turn 6.9s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
❑	d0	c1	412	16.1	5,415	12,960
		c2	402	30.6	10,003	17,612
		c4	407	53.2	20,023	28,248
❑	d4096	c1	415	17.0	15,214	22,329
		c2	411	28.8	29,687	37,570
		c4	404	14.6	52,846	64,504
❑	d8192	c1	396	17.0	26,260	33,376
		c2	397	30.6	51,414	58,864
		c4	400	10.1	84,178	100,848
Per-stream (single request, per concurrent client)						
❑	—	c1	415	17.0	15,214	—
		c2	411	30.6	—	—
		c4	407	53.2	—	—

### 3.2.8 SGLang.runtime + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-4

```

MODEL_PATH="./models/Qwen3.6-27B-INT4-AutoRound-Intel"
MODEL_NAME=$(basename "$MODEL_PATH")
SGLANG_IMAGE="i-dgx-spark-gb10:sglang.runtime"

docker run \
  --rm -it \
  --gpus all \
  --name ${MODEL_NAME} \
  -v ${MODEL_PATH}:/models/${MODEL_NAME} \
  -p 4321:4321 \
  --ipc=host \
  -e SGLANG_ENABLE_SPEC_V2=1 \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  ${SGLANG_IMAGE} python3 -m sglang.launch_server \
    --model-path /models/${MODEL_NAME} \
    --served-model-name ${MODEL_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --mem-fraction-static 0.75 \

```

```

--context-length 262144 \
--max-running-requests 16 \
--cuda-graph-max-bs 16 \
--max-prefill-tokens 16384 \
--chunked-prefill-size 8192 \
--quantization auto-round \
--kv-cache-dtype fp8_e5m2 \
--reasoning-parser qwen3 \
--tool-call-parser qwen3_coder \
--attention-backend flashinfer \
--mamba-scheduler-strategy extra_buffer \
--speculative-algorithm NEXTN \
--speculative-num-steps 4 \
--speculative-eagle-topk 1 \
--speculative-num-draft-tokens 5 \
--trust-remote-code

```

```

HF_HUB_OFFLINE=1 tool-eval-bench \
--base-url http://127.0.0.1:4321 \
--short --perf \
--bench-args "--tokenizer ./models/Qwen3.6-27B-INT4-AutoRound-Intel"

```

GPU 90.14GiB • Warm-up 2537ms • Quality 97/100 • Responsiveness 26/100 • Median turn 6.0s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
🟡	d0	c1	409	22.7	5,479	10,671
		c2	405	32.4	9,944	16,689
		c4	409	58.2	19,898	27,143
🟢	d4096	c1	409	20.3	15,486	21,342
		c2	402	34.2	30,383	36,944
		c4	399	14.8	53,581	64,286
🟢	d8192	c1	396	21.0	26,275	31,917
		c2	394	33.6	51,763	58,352
		c4	397	10.1	84,571	100,418
Per-stream (single request, per concurrent client)						
🟠	—	c1	409	22.7	5,479	—
		c2	405	34.2	—	—
		c4	409	58.2	—	—

### 3.2.9 SGLang.runtime + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-8

```

MODEL_PATH="./models/Qwen3.6-27B-INT4-AutoRound-Intel"
MODEL_NAME=$(basename "$MODEL_PATH")
SGLANG_IMAGE="i-dgx-spark-gb10:sglang.runtime"

docker run \
  --rm -it \
  --gpus all \
  --name ${MODEL_NAME} \
  -v ${MODEL_PATH}:/models/${MODEL_NAME} \
  -p 4321:4321 \
  --ipc=host \
  -e SGLANG_ENABLE_SPEC_V2=1 \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  ${SGLANG_IMAGE} python3 -m sglang.launch_server \
    --model-path /models/${MODEL_NAME} \
    --served-model-name ${MODEL_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --mem-fraction-static 0.75 \
    --context-length 262144 \
    --max-running-requests 16 \
    --cuda-graph-max-bs 16 \
    --max-prefill-tokens 16384 \
    --chunked-prefill-size 8192 \
    --quantization auto-round \
    --kv-cache-dtype fp8_e5m2 \
    --reasoning-parser qwen3 \
    --tool-call-parser qwen3_coder \
    --attention-backend flashinfer \
    --mamba-scheduler-strategy extra_buffer \
    --speculative-algorithm NEXTN \
    --speculative-num-steps 8 \
    --speculative-eagle-topk 1 \
    --speculative-num-draft-tokens 9 \
    --trust-remote-code

HF_HUB_OFFLINE=1 tool-eval-bench \
  --base-url http://127.0.0.1:4321 \
  --short --perf \
  --bench-args "--tokenizer ../models/Qwen3.6-27B-INT4-AutoRound-Intel"

```

GPU 90.94GiB • Warm-up 16813ms • Quality 97/100 • Responsiveness 23/100 • Median turn 6.7s						
	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
❑	d0	c1	411	20.5	5,574	11,288
		c2	394	25.3	10,232	18,605
		c4	346	42.2	23,665	32,676
❑	d4096	c1	404	15.5	15,793	23,500
		c2	380	25.4	32,152	41,046
		c4	361	12.6	59,081	71,013
❑	d8192	c1	359	19.9	29,051	34,940
		c2	372	25.0	54,770	63,292
		c4	378	9.3	89,624	105,330
Per-stream (single request, per concurrent client)						
❑	—	c1	411	20.5	5,574	—
		c2	394	25.4	—	—
		c4	378	42.2	—	—

### 3.2.10 vLLM.dev + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-2

```

MODEL_PATH="./models/Qwen3.6-27B-INT4-AutoRound-Intel"
MODEL_NAME=$(basename "$MODEL_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm0.19.1.dev"

docker run \
  --rm -it \
  --gpus all \
  --name ${MODEL_NAME} \
  -v ${MODEL_PATH}:/models/${MODEL_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${MODEL_NAME} \
    --served-model-name ${MODEL_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --gpu-memory-utilization 0.8 \
    --max-model-len 262144 \

```

```

--max-num-seqs 4 \
--max-num-batched-tokens 16384 \
--speculative-config '{"method": "mtp", "num_speculative_tokens": 2}' \
--reasoning-parser qwen3 \
--tool-call-parser qwen3_coder \
--load-format instanttensor \
--attention-backend flashinfer \
--kv-cache-dtype fp8 \
--quantization gptq_marlin \
--enable-prefix-caching \
--enable-chunked-prefill \
--enable-auto-tool-choice \
--trust-remote-code

```

```

HF_HUB_OFFLINE=1 tool-eval-bench \
--base-url http://127.0.0.1:4321 \
--short --perf \
--bench-args "--tokenizer ./models/Qwen3.6-27B-INT4-AutoRound-Intel"

```

GPU 98.43GiB • Warm-up 3668ms • Quality 97/100 • Responsiveness 17/100 • Median turn 8.8s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
■	d0	c1	453	17.0	4,828	12,070
		c2	410	27.1	9,899	18,465
		c4	401	46.6	20,473	29,273
■	d4096	c1	442	17.5	14,209	21,241
		c2	444	31.5	27,595	35,179
		c4	448	56.0	54,743	63,048
■	d8192	c1	441	16.9	23,515	30,773
		c2	442	30.1	46,308	54,148
		c4	439	25.8	90,431	100,113
Per-stream (single request, per concurrent client)						
■	—	c1	453	17.5	14,209	—
		c2	444	31.5	—	—
		c4	448	56.0	—	—

### 3.2.11 vLLM.dev + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-4

```

MODEL_PATH="./models/Qwen3.6-27B-INT4-AutoRound-Intel"
MODEL_NAME=$(basename "$MODEL_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm0.19.1.dev"

```

```

docker run \
  --rm -it \
  --gpus all \
  --name ${MODEL_NAME} \
  -v ${MODEL_PATH}:/models/${MODEL_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${MODEL_NAME} \
    --served-model-name ${MODEL_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --gpu-memory-utilization 0.8 \
    --max-model-len 262144 \
    --max-num-seqs 4 \
    --max-num-batched-tokens 16384 \
    --speculative-config '{"method": "mtp", "num_speculative_tokens": 4}' \
    --reasoning-parser qwen3 \
    --tool-call-parser qwen3_coder \
    --load-format instanttensor \
    --attention-backend flashinfer \
    --kv-cache-dtype fp8 \
    --quantization gptq_marlin \
    --enable-prefix-caching \
    --enable-chunked-prefill \
    --enable-auto-tool-choice \
    --trust-remote-code

HF_HUB_OFFLINE=1 tool-eval-bench \
  --base-url http://127.0.0.1:4321 \
  --short --perf \
  --bench-args "--tokenizer ./models/Qwen3.6-27B-INT4-AutoRound-Intel"

```



GPU 98.50GiB • Warm-up 3152ms • Quality 97/100 • Responsiveness 18/100 • Median turn 8.4s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
🟦	d0	c1	454	20.2	4,829	10,847
		c2	354	25.2	11,598	20,081
		c4	395	53.9	20,736	27,922
🟪	d4096	c1	443	18.1	14,211	20,957
		c2	442	31.2	27,697	34,944
		c4	450	62.3	54,589	61,714
🟩	d8192	c1	440	18.5	23,603	30,192
		c2	440	36.6	46,459	53,027
		c4	441	26.8	90,115	98,980
Per-stream (single request, per concurrent client)						
🟨	—	c1	454	20.2	4,829	—
		c2	442	36.6	—	—
		c4	450	62.3	—	—

### 3.2.12 vLLM.dev + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-8

```
MODEL_PATH="./models/Qwen3.6-27B-INT4-AutoRound-Intel"
MODEL_NAME=$(basename "$MODEL_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm0.19.1.dev"

docker run \
  --rm -it \
  --gpus all \
  --name ${MODEL_NAME} \
  -v ${MODEL_PATH}:/models/${MODEL_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
  serve /models/${MODEL_NAME} \
  --served-model-name ${MODEL_NAME} \
  --host 0.0.0.0 \
  --port 4321 \
  --gpu-memory-utilization 0.8 \
  --max-model-len 262144 \
```

```

--max-num-seqs 4 \
--max-num-batched-tokens 16384 \
--speculative-config '{"method": "mtp", "num_speculative_tokens": 8}' \
--reasoning-parser qwen3 \
--tool-call-parser qwen3_coder \
--load-format instanttensor \
--attention-backend flashinfer \
--kv-cache-dtype fp8 \
--quantization gptq_marlin \
--enable-prefix-caching \
--enable-chunked-prefill \
--enable-auto-tool-choice \
--trust-remote-code

```

```

HF_HUB_OFFLINE=1 tool-eval-bench \
--base-url http://127.0.0.1:4321 \
--short --perf \
--bench-args "--tokenizer ./models/Qwen3.6-27B-INT4-AutoRound-Intel"

```

GPU **98.50GiB** • Warm-up **2695ms** • Quality **53/100** • Responsiveness **17/100** • Median turn **8.6s**

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
❑	d0	c1	453	19.5	4,909	11,102
		c2	352	26.5	11,792	19,951
		c4	447	42.6	18,235	27,472
❑	d4096	c1	431	14.2	14,640	23,301
		c2	443	31.2	27,630	34,789
		c4	448	44.6	54,759	63,419
❑	d8192	c1	438	17.3	23,746	30,755
		c2	436	30.5	46,809	54,134
		c4	440	24.5	90,408	100,781
Per-stream (single request, per concurrent client)						
❑	—	c1	453	19.5	4,909	—
		c2	443	31.2	—	—
		c4	448	44.6	—	—

### 3.2.13 vLLM.build + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-2

```

MODEL_PATH="./models/Qwen3.6-27B-INT4-AutoRound-Intel"
MODEL_NAME=$(basename "$MODEL_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm0.20.0.build"

```

```

docker run \
  --rm -it \
  --gpus all \
  --name ${MODEL_NAME} \
  -v ${MODEL_PATH}:/models/${MODEL_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${MODEL_NAME} \
    --served-model-name ${MODEL_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --gpu-memory-utilization 0.8 \
    --max-model-len 262144 \
    --max-num-seqs 4 \
    --max-num-batched-tokens 16384 \
    --speculative-config '{"method": "mtp", "num_speculative_tokens": 2}' \
    --reasoning-parser qwen3 \
    --tool-call-parser qwen3_coder \
    --load-format instanttensor \
    --attention-backend flashinfer \
    --kv-cache-dtype fp8 \
    --quantization gptq_marlin \
    --enable-prefix-caching \
    --enable-chunked-prefill \
    --enable-auto-tool-choice \
    --trust-remote-code

HF_HUB_OFFLINE=1 tool-eval-bench \
  --base-url http://127.0.0.1:4321 \
  --short --perf \
  --bench-args "--tokenizer ./models/Qwen3.6-27B-INT4-AutoRound-Intel"

```

GPU 98.19GiB • Warm-up 599ms • Quality 96/100 • Responsiveness 14/100 • Median turn 10.1s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
🟦	d0	c1	429	16.0	5,054	12,789
		c2	389	29.4	10,411	18,309
		c4	375	48.6	21,837	30,146
🟪	d4096	c1	425	17.1	14,746	21,955
		c2	423	30.8	28,942	36,659
		c4	425	57.8	57,733	65,785
🟩	d8192	c1	417	16.5	24,814	32,301
		c2	417	29.5	49,062	56,968
		c4	418	24.9	95,065	104,683
Per-stream (single request, per concurrent client)						
🟨	—	c1	429	17.1	14,746	—
		c2	423	30.8	—	—
		c4	425	57.8	—	—

### 3.2.14 vLLM.build + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-4

```

MODEL_PATH="./models/Qwen3.6-27B-INT4-AutoRound-Intel"
MODEL_NAME=$(basename "$MODEL_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm0.20.0.build"

docker run \
  --rm -it \
  --gpus all \
  --name ${MODEL_NAME} \
  -v ${MODEL_PATH}:/models/${MODEL_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${MODEL_NAME} \
    --served-model-name ${MODEL_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --gpu-memory-utilization 0.8 \
    --max-model-len 262144 \

```

```

--max-num-seqs 4 \
--max-num-batched-tokens 16384 \
--speculative-config '{"method": "mtp", "num_speculative_tokens": 4}' \
--reasoning-parser qwen3 \
--tool-call-parser qwen3_coder \
--load-format instanttensor \
--attention-backend flashinfer \
--kv-cache-dtype fp8 \
--quantization gptq_marlin \
--enable-prefix-caching \
--enable-chunked-prefill \
--enable-auto-tool-choice \
--trust-remote-code

```

```

HF_HUB_OFFLINE=1 tool-eval-bench \
--base-url http://127.0.0.1:4321 \
--short --perf \
--bench-args "--tokenizer ./models/Qwen3.6-27B-INT4-AutoRound-Intel"

```

GPU 100.30GiB • Warm-up 16820ms • Quality 97/100 • Responsiveness 31/100 • Median turn 5.3s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
■	d0	c1	980	27.1	2,327	6,816
		c2	913	51.2	4,442	8,831
		c4	954	91.3	8,554	13,126
■	d4096	c1	969	24.0	6,575	11,676
		c2	949	53.4	12,878	17,059
		c4	899	81.5	24,662	29,437
■	d8192	c1	926	28.7	11,299	15,523
		c2	919	39.0	20,354	25,765
		c4	854	20.6	35,126	41,546
Per-stream (single request, per concurrent client)						
■	—	c1	980	28.7	11,299	—
		c2	949	53.4	—	—
		c4	954	91.3	—	—

### 3.2.15 vLLM.build + Qwen3.6-27B-INT4-AutoRound-Intel + MTP-8

```

MODEL_PATH="./models/Qwen3.6-27B-INT4-AutoRound-Intel"
MODEL_NAME=$(basename "$MODEL_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm0.20.0.build"

```

```

docker run \
  --rm -it \
  --gpus all \
  --name ${MODEL_NAME} \
  -v ${MODEL_PATH}:/models/${MODEL_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${MODEL_NAME} \
    --served-model-name ${MODEL_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --gpu-memory-utilization 0.8 \
    --max-model-len 262144 \
    --max-num-seqs 4 \
    --max-num-batched-tokens 16384 \
    --speculative-config '{"method": "mtp", "num_speculative_tokens": 8}' \
    --reasoning-parser qwen3 \
    --tool-call-parser qwen3_coder \
    --load-format instanttensor \
    --attention-backend flashinfer \
    --kv-cache-dtype fp8 \
    --quantization gptq_marlin \
    --enable-prefix-caching \
    --enable-chunked-prefill \
    --enable-auto-tool-choice \
    --trust-remote-code

HF_HUB_OFFLINE=1 tool-eval-bench \
  --base-url http://127.0.0.1:4321 \
  --short --perf \
  --bench-args "--tokenizer ./models/Qwen3.6-27B-INT4-AutoRound-Intel"

```

GPU 98.75GiB • Warm-up 15674ms • Quality 53/100 • Responsiveness 25/100 • Median turn 6.1s						
	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
❑	d0	c1	912	21.7	2,525	8,145
		c2	914	46.0	4,452	9,360
		c4	933	69.4	8,715	14,494
❑	d4096	c1	956	26.9	6,704	11,181
		c2	942	46.1	12,970	17,461
		c4	962	77.3	25,470	30,537
❑	d8192	c1	953	22.5	11,015	16,438
		c2	943	46.7	21,622	26,333
		c4	953	50.6	41,748	47,153
Per-stream (single request, per concurrent client)						
❑	—	c1	956	26.9	6,704	—
		c2	943	46.7	—	—
		c4	962	77.3	—	—

### 3.2.16 vLLM.build + Qwen3.6-27B-INT4-AutoRound-Intel + DFlash-2

```

TARGET_PATH="./models/Qwen3.6-27B-INT4-AutoRound-Intel"
DRAFT_PATH="./models/Qwen3.6-27B-DFlash"
TARGET_NAME=$(basename "$TARGET_PATH")
DRAFT_NAME=$(basename "$DRAFT_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm0.20.0.build"

docker run \
  --rm -it \
  --gpus all \
  --name ${TARGET_NAME}-dflash \
  -v $(pwd)/${TARGET_PATH}:/models/${TARGET_NAME} \
  -v $(pwd)/${DRAFT_PATH}:/models/${DRAFT_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${TARGET_NAME} \
    --served-model-name ${TARGET_NAME} \
    --host 0.0.0.0 \

```

```

--port 4321 \
--gpu-memory-utilization 0.8 \
--max-model-len 262144 \
--max-num-seqs 4 \
--max-num-batched-tokens 16384 \
--speculative-config '{"model": "/models/'${DRAFT_NAME}'", "method": "dflash",
"num_speculative_tokens": 3}' \
--reasoning-parser qwen3 \
--tool-call-parser qwen3_coder \
--attention-backend FLASH_ATTN \
--enable-prefix-caching \
--enable-chunked-prefill \
--enable-auto-tool-choice \
--trust-remote-code

```

```

HF_HUB_OFFLINE=1 tool-eval-bench \
--base-url http://127.0.0.1:4321 \
--short --perf \
--bench-yargs "--tokenizer ./models/Qwen3.6-27B-INT4-AutoRound-Intel"

```

GPU 96.91GiB • Warm-up 32629ms • Quality 43/100 • Responsiveness 15/100 • Median turn 9.3s						
Depth		Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
<i>Aggregate throughput (pp2048 / tg128)</i>						
□	d0	c1	990	28.1	<b>2,312</b>	<b>6,629</b>
		c2	901	48.0	4,522	9,387
		c4	885	79.6	9,188	14,221
□	d4096	c1	<b>1,026</b>	25.0	6,224	11,103
		c2	998	46.1	12,254	17,202
		c4	1,013	<b>79.7</b>	24,226	29,741
□	d8192	c1	1,014	23.4	10,332	15,557
		c2	1,005	42.1	20,325	25,950
		c4	992	49.6	40,530	48,015
<i>Per-stream (single request, per concurrent client)</i>						
□	—	c1	1,026	28.1	2,312	—
		c2	1,005	48.0	—	—
		c4	1,013	79.7	—	—

### 3.2.17 vLLM.build + Qwen3.6-27B-INT4-AutoRound-Intel + DFlash-4

```

TARGET_PATH="./models/Qwen3.6-27B-INT4-AutoRound-Intel"
DRAFT_PATH="./models/Qwen3.6-27B-DFlash"

```



```

TARGET_NAME=$(basename "$TARGET_PATH")
DRAFT_NAME=$(basename "$DRAFT_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm0.20.0.build"

docker run \
  --rm -it \
  --gpus all \
  --name ${TARGET_NAME}-dflash \
  -v $(pwd)/${TARGET_PATH}:/models/${TARGET_NAME} \
  -v $(pwd)/${DRAFT_PATH}:/models/${DRAFT_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${TARGET_NAME} \
    --served-model-name ${TARGET_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --gpu-memory-utilization 0.8 \
    --max-model-len 262144 \
    --max-num-seqs 4 \
    --max-num-batched-tokens 16384 \
    --speculative-config '{"model": "/models/'${DRAFT_NAME}', "method": "dflash",
"num_speculative_tokens": 4}' \
    --reasoning-parser qwen3 \
    --tool-call-parser qwen3_coder \
    --attention-backend FLASH_ATTN \
    --enable-prefix-caching \
    --enable-chunked-prefill \
    --enable-auto-tool-choice \
    --trust-remote-code

HF_HUB_OFFLINE=1 tool-eval-bench \
  --base-url http://127.0.0.1:4321 \
  --short --perf \
  --bench-args "--tokenizer ./models/Qwen3.6-27B-INT4-AutoRound-Intel"

```

GPU 95.42GiB • Warm-up 32549ms • Quality 43/100 • Responsiveness 17/100 • Median turn 8.6s						
	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
❑	d0	c1	951	32.7	2,382	6,084
		c2	792	47.0	5,101	9,782
		c4	885	77.6	9,170	14,055
❑	d4096	c1	1,023	26.1	6,214	10,903
		c2	1,004	49.9	12,184	16,888
		c4	1,002	84.1	24,470	29,757
❑	d8192	c1	1,015	23.7	10,302	15,491
		c2	1,003	40.1	20,368	25,802
		c4	999	50.3	40,238	47,372
Per-stream (single request, per concurrent client)						
❑	—	c1	1,023	32.7	2,382	—
		c2	1004	49.9	—	—
		c4	1,002	84.1	—	—

### 3.2.18 vLLM.build + Qwen3.6-27B-INT4-AutoRound-Intel + DFlash-8

```

TARGET_PATH="./models/Qwen3.6-27B-INT4-AutoRound-Intel"
DRAFT_PATH="./models/Qwen3.6-27B-DFlash"
TARGET_NAME=$(basename "$TARGET_PATH")
DRAFT_NAME=$(basename "$DRAFT_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm0.20.0.build"

docker run \
  --rm -it \
  --gpus all \
  --name ${TARGET_NAME}-dflash \
  -v $(pwd)/${TARGET_PATH}:/models/${TARGET_NAME} \
  -v $(pwd)/${DRAFT_PATH}:/models/${DRAFT_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${TARGET_NAME} \
    --served-model-name ${TARGET_NAME} \
    --host 0.0.0.0 \

```

```

--port 4321 \
--gpu-memory-utilization 0.8 \
--max-model-len 262144 \
--max-num-seqs 4 \
--max-num-batched-tokens 16384 \
--speculative-config '{"model": "/models/'${DRAFT_NAME}'", "method": "dflash",
"num_speculative_tokens": 8}' \
--reasoning-parser qwen3 \
--tool-call-parser qwen3_coder \
--attention-backend FLASH_ATTN \
--enable-prefix-caching \
--enable-chunked-prefill \
--enable-auto-tool-choice \
--trust-remote-code

```

```

HF_HUB_OFFLINE=1 tool-eval-bench \
--base-url http://127.0.0.1:4321 \
--short --perf \
--bench-args "--tokenizer ./models/Qwen3.6-27B-INT4-AutoRound-Intel"

```

GPU 95.28GiB • Warm-up 587ms • Quality 43/100 • Responsiveness 8/100 • Median turn 15.5s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
<i>Aggregate throughput (pp2048 / tg128)</i>						
□	d0	c1	406	20.6	<b>5,351</b>	<b>11,309</b>
		c2	344	30.1	11,858	18,836
		c4	383	44.7	21,305	28,982
□	d4096	c1	<b>429</b>	17.5	14,577	21,619
		c2	<b>429</b>	29.8	28,569	36,546
		c4	421	<b>48.3</b>	58,246	67,266
□	d8192	c1	<b>429</b>	14.5	24,120	32,707
		c2	426	25.0	47,920	57,321
		c4	418	27.0	96,609	109,596
<i>Per-stream (single request, per concurrent client)</i>						
□	—	c1	429	20.6	5,351	—
		c2	429	30.1	—	—
		c4	421	48.3	—	—

### 3.2.19 vLLM.build + Qwen3.6-27B-INT4-AutoRound-Intel + DFlash-15

```

TARGET_PATH="./models/Qwen3.6-27B-INT4-AutoRound-Intel"
DRAFT_PATH="./models/Qwen3.6-27B-DFlash"

```

```

TARGET_NAME=$(basename "$TARGET_PATH")
DRAFT_NAME=$(basename "$DRAFT_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm0.20.0.build"

docker run \
  --rm -it \
  --gpus all \
  --name ${TARGET_NAME}-dflash \
  -v $(pwd)/${TARGET_PATH}:/models/${TARGET_NAME} \
  -v $(pwd)/${DRAFT_PATH}:/models/${DRAFT_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${TARGET_NAME} \
    --served-model-name ${TARGET_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --gpu-memory-utilization 0.8 \
    --max-model-len 262144 \
    --max-num-seqs 4 \
    --max-num-batched-tokens 16384 \
    --speculative-config '{"model": "/models/'${DRAFT_NAME}', "method": "dflash",
    "num_speculative_tokens": 15}' \
    --reasoning-parser qwen3 \
    --tool-call-parser qwen3_coder \
    --attention-backend FLASH_ATTN \
    --enable-prefix-caching \
    --enable-chunked-prefill \
    --enable-auto-tool-choice \
    --trust-remote-code

HF_HUB_OFFLINE=1 tool-eval-bench \
  --base-url http://127.0.0.1:4321 \
  --short --perf \
  --bench-args "--tokenizer ./models/Qwen3.6-27B-INT4-AutoRound-Intel"

```

GPU 95.46GiB • Warm-up 32836ms • Quality 47/100 • Responsiveness 23/100 • Median turn 6.8s						
Depth		Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
❑	d0	c1	970	35.3	2,332	5,742
		c2	877	61.1	4,621	8,030
		c4	962	78.9	8,449	13,626
❑	d4096	c1	1,016	30.5	6,263	10,249
		c2	993	42.4	12,300	17,299
		c4	992	56.5	24,336	31,500
❑	d8192	c1	1,004	20.7	10,412	16,376
		c2	989	37.7	20,642	26,947
		c4	992	37.8	40,316	49,430
Per-stream (single request, per concurrent client)						
❑	—	c1	1,016	35.3	2,332	—
		c2	993	61.1	—	—
		c4	992	78.9	—	—

### 3.2.20 vLLM.dflash + Qwen3.6-27B-INT4-AutoRound-Intel + DFlash-15

```

TARGET_PATH="./models/Qwen3.6-27B-INT4-AutoRound-Intel"
DRAFT_PATH="./models/Qwen3.6-27B-DFlash"
TARGET_NAME=$(basename "$TARGET_PATH")
DRAFT_NAME=$(basename "$DRAFT_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm.dflash"

docker run \
  --rm -it \
  --gpus all \
  --name ${TARGET_NAME}-dflash \
  -v $(pwd)/${TARGET_PATH}:/models/${TARGET_NAME} \
  -v $(pwd)/${DRAFT_PATH}:/models/${DRAFT_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${TARGET_NAME} \
    --served-model-name ${TARGET_NAME} \
    --host 0.0.0.0 \

```

```

--port 4321 \
--gpu-memory-utilization 0.8 \
--max-model-len 262144 \
--max-num-seqs 4 \
--max-num-batched-tokens 16384 \
--speculative-config '{"model": "/models/'${DRAFT_NAME}'", "method": "dflash",
"num_speculative_tokens": 15}' \
--reasoning-parser qwen3 \
--tool-call-parser qwen3_coder \
--attention-backend FLASH_ATTN \
--enable-prefix-caching \
--enable-chunked-prefill \
--enable-auto-tool-choice \
--trust-remote-code

```

```

HF_HUB_OFFLINE=1 tool-eval-bench \
--base-url http://127.0.0.1:4321 \
--short --perf \
--bench-args "--tokenizer ./models/Qwen3.6-27B-INT4-AutoRound-Intel"

```

GPU 97.02GiB • Warm-up 49539ms • Quality 97/100 • Responsiveness 42/100 • Median turn 3.8s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
<i>Aggregate throughput (pp2048 / tg128)</i>						
■	d0	c1	974	53.9	<b>2,353</b>	<b>4,484</b>
		c2	854	56.2	4,777	8,621
		c4	972	<b>79.0</b>	8,359	13,384
■	d4096	c1	<b>1,012</b>	34.0	6,316	9,844
		c2	992	69.7	12,320	15,483
		c4	992	73.0	24,636	29,898
■	d8192	c1	1,009	34.7	10,392	13,842
		c2	994	66.9	20,531	23,775
		c4	988	66.8	40,917	45,994
<i>Per-stream (single request, per concurrent client)</i>						
■	—	c1	1,012	53.9	2,353	—
		c2	994	69.7	—	—
		c4	992	79.0	—	—

### 3.2.21 LLaMA.cpp + Qwen3.6-27B-Q4\_K\_M + MTP-2

```

llama-server \
-m ./models/Qwen3.6-27B-Q4_K_M.gguf \

```

```

--alias Qwen3.6-27B-Q4_K_M.gguf \
--host 127.0.0.1 \
--port 4321 \
--ngl 99 \
--c 262144 \
--b 8192 \
--ub 4096 \
--np 4 \
--t -1 \
--cache-ram 0 \
--cache-reuse 512 \
--fa on \
--reasoning on \
--spec-type draft-mtp \
--spec-draft-n-max 2 \
--jinja

HF_HUB_OFFLINE=1 tool-eval-bench \
--base-url http://127.0.0.1:4321 \
--short --perf \
--bench-args "--tokenizer ./models/Qwen3.6-27B"

```

GPU 37.96GiB • Warm-up 676ms • Quality 90/100 • Responsiveness 25/100 • Median turn 6.1s						
Depth		Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
🟡	d0	c1	807	18.3	2,805	9,547
		c2	737	19.9	5,412	17,774
		c4	750	23.2	10,736	31,544
🟢	d4096	c1	797	21.4	7,978	13,683
		c2	769	19.9	15,832	28,111
		c4	768	23.1	31,736	52,586
🟢	d8192	c1	754	20.3	13,854	19,896
		c2	739	20.2	27,567	39,709
		c4	729	14.4	50,000	74,823
Per-stream (single request, per concurrent client)						
🟠	—	c1	807	21.4	7,978	—
		c2	769	20.2	—	—
		c4	768	23.2	—	—

### 3.2.22 LLaMA.cpp + Qwen3.6-27B-Q4\_K\_M + MTP-4

```

llama-server \
  -m ./models/Qwen3.6-27B-Q4_K_M.gguf \
  --alias Qwen3.6-27B-Q4_K_M.gguf \
  --host 127.0.0.1 \
  --port 4321 \
  -ngl 99 \
  -c 262144 \
  -b 8192 \
  -ub 4096 \
  -np 4 \
  -t -1 \
  --cache-ram 0 \
  --cache-reuse 512 \
  -fa on \
  --reasoning on \
  --spec-type draft-mtp \
  --spec-draft-n-max 4 \
  --jinja

HF_HUB_OFFLINE=1 tool-eval-bench \
  --base-url http://127.0.0.1:4321 \
  --short --perf \
  --bench-args "--tokenizer ./models/Qwen3.6-27B"

```

GPU **39.13GiB** • Warm-up **501ms** • Quality **93/100** • Responsiveness **29/100** • Median turn **5.5s**

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
🟡	d0	c1	794	24.6	2,807	7,783
		c2	722	22.9	5,503	15,739
		c4	729	26.2	10,958	28,317
🟢	d4096	c1	785	21.3	8,058	13,847
		c2	756	25.1	16,070	25,208
		c4	762	28.2	32,099	48,843
🟢	d8192	c1	750	21.8	13,885	19,529
		c2	735	26.6	27,680	36,773
		c4	723	16.5	50,517	70,598
Per-stream (single request, per concurrent client)						
🟠	—	c1	794	24.6	2,807	—
		c2	756	26.6	—	—
		c4	762	28.2	—	—



### 3.2.23 LLaMA.cpp + Qwen3.6-27B-Q4\_K\_M + MTP-8

```
llama-server \  
  -m ./models/Qwen3.6-27B-Q4_K_M.gguf \  
  --alias Qwen3.6-27B-Q4_K_M.gguf \  
  --host 127.0.0.1 \  
  --port 4321 \  
  -ngl 99 \  
  -c 262144 \  
  -b 8192 \  
  -ub 4096 \  
  -np 4 \  
  -t -1 \  
  --cache-ram 0 \  
  --cache-reuse 512 \  
  -fa on \  
  --reasoning on \  
  --spec-type draft-mtp \  
  --spec-draft-n-max 8 \  
  --jinja  
  
HF_HUB_OFFLINE=1 tool-eval-bench \  
  --base-url http://127.0.0.1:4321 \  
  --short --perf \  
  --bench-args "--tokenizer ./models/Qwen3.6-27B"
```

GPU 41.47GiB • Warm-up 785ms • Quality 100/100 • Responsiveness 27/100 • Median turn 5.8s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
❑	d0	c1	797	19.6	2,828	9,088
		c2	707	18.6	5,586	18,295
		c4	729	26.1	11,072	27,639
❑	d4096	c1	787	27.3	8,063	12,494
		c2	748	25.5	16,207	25,502
		c4	756	27.4	32,295	48,293
❑	d8192	c1	749	23.1	13,929	19,206
		c2	731	20.8	27,826	39,236
		c4	720	18.0	52,446	70,228
Per-stream (single request, per concurrent client)						
❑	—	c1	797	27.3	8,063	—
		c2	748	25.5	—	—
		c4	756	27.4	—	—

### 3.3 Qwen3.5-122B

#### 3.3.1 vLLM.dev + Qwen3.5-122B-A10B-INT4-AutoRound-Intel + MTP-2

```

MODEL_PATH="./models/Qwen3.5-122B-A10B-INT4-AutoRound-Intel"
MODEL_NAME=$(basename "$MODEL_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm0.19.1.dev"

docker run \
  --rm -it \
  --gpus all \
  --name ${MODEL_NAME} \
  -v ${MODEL_PATH}:/models/${MODEL_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${MODEL_NAME} \
    --served-model-name ${MODEL_NAME} \
    --host 0.0.0.0 \
    --port 4321 \

```

```

--gpu-memory-utilization 0.8 \
--max-model-len 262144 \
--max-num-seqs 4 \
--max-num-batched-tokens 16384 \
--speculative-config '{"method": "mtp", "num_speculative_tokens": 2}' \
--reasoning-parser qwen3 \
--tool-call-parser qwen3_coder \
--load-format instanttensor \
--attention-backend flashinfer \
--kv-cache-dtype fp8 \
--quantization gptq_marlin \
--enable-prefix-caching \
--enable-chunked-prefill \
--enable-auto-tool-choice \
--trust-remote-code

```

```

HF_HUB_OFFLINE=1 tool-eval-bench \
--base-url http://127.0.0.1:4321 \
--short --perf \
--bench-args "--tokenizer ./models/Qwen3.5-122B-A10B-INT4-AutoRound-Intel"

```

GPU 98.88GiB • Warm-up 12355ms • Quality 100/100 • Responsiveness 48/100 • Median turn 3.1s						
Depth		Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
❑	d0	c1	1,457	39.4	2,120	4,864
		c2	1,435	59.4	2,941	6,510
		c4	1,675	79.7	5,178	10,152
❑	d4096	c1	2,421	38.1	3,048	5,904
		c2	2,162	59.7	5,625	9,200
		c4	2,205	74.4	10,827	16,069
❑	d8192	c1	2,329	39.3	4,900	7,654
		c2	2,195	61.4	9,303	12,912
		c4	2,217	58.2	17,696	23,413
Per-stream (single request, per concurrent client)						
❑	—	c1	2,421	39.4	2,120	—
		c2	2,195	61.4	—	—
		c4	2,217	79.7	—	—

### 3.3.2 vLLM.dev + Qwen3.5-122B-A10B-INT4-AutoRound-Intel + MTP-4

```
MODEL_PATH="./models/Qwen3.5-122B-A10B-INT4-AutoRound-Intel"
```

```

MODEL_NAME=$(basename "$MODEL_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm0.19.1.dev"

docker run \
  --rm -it \
  --gpus all \
  --name ${MODEL_NAME} \
  -v ${MODEL_PATH}:/models/${MODEL_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${MODEL_NAME} \
    --served-model-name ${MODEL_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --gpu-memory-utilization 0.8 \
    --max-model-len 262144 \
    --max-num-seqs 4 \
    --max-num-batched-tokens 16384 \
    --speculative-config '{"method": "mtp", "num_speculative_tokens": 4}' \
    --reasoning-parser qwen3 \
    --tool-call-parser qwen3_coder \
    --load-format instanttensor \
    --attention-backend flashinfer \
    --kv-cache-dtype fp8 \
    --quantization gptq_marlin \
    --enable-prefix-caching \
    --enable-chunked-prefill \
    --enable-auto-tool-choice \
    --trust-remote-code

HF_HUB_OFFLINE=1 tool-eval-bench \
  --base-url http://127.0.0.1:4321 \
  --short --perf \
  --bench-args "--tokenizer ./models/Qwen3.5-122B-A10B-INT4-AutoRound-Intel"

```

GPU 99.32GiB • Warm-up 1366ms • Quality 97/100 • Responsiveness 46/100 • Median turn 3.3s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
🟡	d0	c1	1,896	39.1	1,348	4,383
		c2	1,407	50.8	3,087	7,299
		c4	1,753	74.1	4,782	10,346
🟢	d4096	c1	2,333	37.5	2,870	6,046
		c2	2,168	53.7	5,608	9,835
		c4	2,160	70.5	11,009	16,894
🟢	d8192	c1	2,219	38.6	4,851	7,934
		c2	2,200	59.6	9,244	13,077
		c4	2,195	56.5	17,846	23,926
Per-stream (single request, per concurrent client)						
🟡	—	c1	2,333	39.1	1,348	—
		c2	2,200	59.6	—	—
		c4	2,195	74.1	—	—

### 3.3.3 vLLM.dev + Qwen3.5-122B-A10B-INT4-AutoRound-Intel + MTP-8

```

MODEL_PATH="./models/Qwen3.5-122B-A10B-INT4-AutoRound-Intel"
MODEL_NAME=$(basename "$MODEL_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm0.19.1.dev"

docker run \
  --rm -it \
  --gpus all \
  --name ${MODEL_NAME} \
  -v ${MODEL_PATH}:/models/${MODEL_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${MODEL_NAME} \
    --served-model-name ${MODEL_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --gpu-memory-utilization 0.8 \
    --max-model-len 262144 \

```

```

--max-num-seqs 4 \
--max-num-batched-tokens 16384 \
--speculative-config '{"method": "mtp", "num_speculative_tokens": 8}' \
--reasoning-parser qwen3 \
--tool-call-parser qwen3_coder \
--load-format instanttensor \
--attention-backend flashinfer \
--kv-cache-dtype fp8 \
--quantization gptq_marlin \
--enable-prefix-caching \
--enable-chunked-prefill \
--enable-auto-tool-choice \
--trust-remote-code

```

```

HF_HUB_OFFLINE=1 tool-eval-bench \
--base-url http://127.0.0.1:4321 \
--short --perf \
--bench-args "--tokenizer ./models/Qwen3.5-122B-A10B-INT4-AutoRound-Intel"

```

GPU 99.13GiB • Warm-up 1620ms • Quality 73/100 • Responsiveness 40/100 • Median turn 4.0s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
❑	d0	c1	1,813	33.0	1,435	5,041
		c2	1,243	47.7	3,848	8,582
		c4	2,027	75.4	3,929	9,958
❑	d4096	c1	2,327	27.4	2,915	7,315
		c2	2,151	44.5	5,622	10,670
		c4	2,155	58.0	10,950	17,532
❑	d8192	c1	2,157	26.3	5,024	9,610
		c2	2,155	42.6	9,410	14,814
		c4	2,180	49.3	17,893	24,879
Per-stream (single request, per concurrent client)						
❑	—	c1	2,327	33.0	1,435	—
		c2	2,155	47.7	—	—
		c4	2,180	75.4	—	—

### 3.3.4 vLLM.dflash + Qwen3.5-122B-A10B-INT4-AutoRound-Intel + DFlash-2

```

TARGET_PATH="./models/Qwen3.5-122B-A10B-INT4-AutoRound-Intel"
DRAFT_PATH="./models/Qwen3.5-122B-A10B-DFlash"
TARGET_NAME=$(basename "$TARGET_PATH")

```

```

DRAFT_NAME=$(basename "$DRAFT_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm.dflash"

docker run \
  --rm -it \
  --gpus all \
  --name ${TARGET_NAME}-dflash \
  -v $(pwd)/${TARGET_PATH}:/models/${TARGET_NAME} \
  -v $(pwd)/${DRAFT_PATH}:/models/${DRAFT_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${TARGET_NAME} \
    --served-model-name ${TARGET_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --gpu-memory-utilization 0.8 \
    --max-model-len 262144 \
    --max-num-seqs 4 \
    --max-num-batched-tokens 16384 \
    --speculative-config '{"model": "/models/'${DRAFT_NAME}'", "method": "dflash",
    "num_speculative_tokens": 2}' \
    --reasoning-parser qwen3 \
    --tool-call-parser qwen3_coder \
    --attention-backend FLASH_ATTN \
    --enable-prefix-caching \
    --enable-chunked-prefill \
    --enable-auto-tool-choice \
    --trust-remote-code

HF_HUB_OFFLINE=1 tool-eval-bench \
  --base-url http://127.0.0.1:4321 \
  --short --perf \
  --bench-args "--tokenizer ./models/Qwen3.5-122B-A10B-INT4-AutoRound-Intel"

```

GPU 98.93GiB • Warm-up 50594ms • Quality 97/100 • Responsiveness 48/100 • Median turn 3.1s						
	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
❑	d0	c1	1,631	42.5	1,544	4,315
		c2	1,636	63.5	2,568	6,193
		c4	1,787	66.7	4,714	11,261
❑	d4096	c1	2,185	40.9	3,059	5,951
		c2	2,187	57.5	5,586	9,653
		c4	2,315	69.4	10,556	17,492
❑	d8192	c1	2,306	39.5	4,680	7,678
		c2	2,278	58.1	8,943	12,987
		c4	2,301	54.5	17,265	24,703
Per-stream (single request, per concurrent client)						
❑	—	c1	2,306	42.5	1,544	—
		c2	2,278	63.5	—	—
		c4	2,315	69.4	—	—

### 3.3.5 vLLM.dflash + Qwen3.5-122B-A10B-INT4-AutoRound-Intel + DFlash-8

```

TARGET_PATH="./models/Qwen3.5-122B-A10B-INT4-AutoRound-Intel"
DRAFT_PATH="./models/Qwen3.5-122B-A10B-DFlash"
TARGET_NAME=$(basename "$TARGET_PATH")
DRAFT_NAME=$(basename "$DRAFT_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm.dflash"

docker run \
  --rm -it \
  --gpus all \
  --name ${TARGET_NAME}-dflash \
  -v $(pwd)/${TARGET_PATH}:/models/${TARGET_NAME} \
  -v $(pwd)/${DRAFT_PATH}:/models/${DRAFT_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${TARGET_NAME} \
    --served-model-name ${TARGET_NAME} \
    --host 0.0.0.0 \

```



```

--port 4321 \
--gpu-memory-utilization 0.8 \
--max-model-len 262144 \
--max-num-seqs 4 \
--max-num-batched-tokens 16384 \
--speculative-config '{"model": "/models/'${DRAFT_NAME}'", "method": "dflash",
"num_speculative_tokens": 8}' \
--reasoning-parser qwen3 \
--tool-call-parser qwen3_coder \
--attention-backend FLASH_ATTN \
--enable-prefix-caching \
--enable-chunked-prefill \
--enable-auto-tool-choice \
--trust-remote-code

```

```

HF_HUB_OFFLINE=1 tool-eval-bench \
--base-url http://127.0.0.1:4321 \
--short --perf \
--bench-args "--tokenizer ./models/Qwen3.5-122B-A10B-INT4-AutoRound-Intel"

```

GPU 100.20GiB • Warm-up 52185ms • Quality 100/100 • Responsiveness 46/100 • Median turn 3.3s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
□	d0	c1	2,279	52.5	1,128	3,335
		c2	1,860	63.8	2,150	5,628
		c4	2,176	91.7	3,707	8,533
□	d4096	c1	2,245	49.1	2,965	5,345
		c2	2,178	54.3	5,593	9,608
		c4	2,336	80.0	10,481	16,184
□	d8192	c1	2,316	44.8	4,650	7,278
		c2	2,093	43.2	9,768	15,021
		c4	2,090	52.1	17,098	23,500
Per-stream (single request, per concurrent client)						
□	—	c1	2,306	51.9	1,660	—
		c2	2,281	60.4	—	—
		c4	2,304	95.3	—	—

### 3.3.6 vLLM.dflash + Qwen3.5-122B-A10B-INT4-AutoRound-Intel + DFlash-15

```

TARGET_PATH="./models/Qwen3.5-122B-A10B-INT4-AutoRound-Intel"
DRAFT_PATH="./models/Qwen3.5-122B-A10B-DFlash"

```

```

TARGET_NAME=$(basename "$TARGET_PATH")
DRAFT_NAME=$(basename "$DRAFT_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm.dflash"

docker run \
  --rm -it \
  --gpus all \
  --name ${TARGET_NAME}-dflash \
  -v $(pwd)/${TARGET_PATH}:/models/${TARGET_NAME} \
  -v $(pwd)/${DRAFT_PATH}:/models/${DRAFT_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${TARGET_NAME} \
    --served-model-name ${TARGET_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --gpu-memory-utilization 0.8 \
    --max-model-len 262144 \
    --max-num-seqs 4 \
    --max-num-batched-tokens 16384 \
    --speculative-config '{"model": "/models/'${DRAFT_NAME}'", "method": "dflash",
    "num_speculative_tokens": 15}' \
    --reasoning-parser qwen3 \
    --tool-call-parser qwen3_coder \
    --attention-backend FLASH_ATTN \
    --enable-prefix-caching \
    --enable-chunked-prefill \
    --enable-auto-tool-choice \
    --trust-remote-code

HF_HUB_OFFLINE=1 tool-eval-bench \
  --base-url http://127.0.0.1:4321 \
  --short --perf \
  --bench-args "--tokenizer ./models/Qwen3.5-122B-A10B-INT4-AutoRound-Intel"

```

GPU 100.10GiB • Warm-up 55349ms • Quality 100/100 • Responsiveness 40/100 • Median turn 3.9s						
	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
🟡	d0	c1	2,131	35.9	1,155	4,529
		c2	1,782	55.9	2,228	6,245
		c4	2,060	73.4	3,908	10,146
🟢	d4096	c1	2,180	39.3	3,013	6,072
		c2	2,149	50.5	5,641	10,288
		c4	2,267	68.6	10,778	17,259
🟢	d8192	c1	2,248	36.9	4,749	8,028
		c2	2,250	48.7	9,025	13,703
		c4	2,231	57.4	17,936	25,038
Per-stream (single request, per concurrent client)						
🟠	—	c1	2,248	39.3	3,013	—
		c2	2,250	55.9	—	—
		c4	2,267	73.4	—	—

### 3.3.7 vLLM.dev + Qwen3.5-122B-Hybrid-INT4FP8 + MTP-2

```

HYBRID_PATH="./models/Qwen3.5-122B-Hybrid-INT4FP8"
HYBRID_NAME=$(basename "$HYBRID_PATH")

VLLM_IMAGE="i-dgx-spark-gb10:vllm0.19.1.dev"

docker run \
  --rm -it \
  --gpus all \
  --name ${HYBRID_NAME} \
  -v $(pwd)/${HYBRID_PATH}:/models/${HYBRID_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${HYBRID_NAME} \
    --served-model-name ${HYBRID_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --gpu-memory-utilization 0.8 \

```

```

--max-model-len 262144 \
--max-num-seqs 4 \
--max-num-batched-tokens 16384 \
--speculative-config '{"method": "mtp", "num_speculative_tokens": 2}' \
--reasoning-parser qwen3 \
--tool-call-parser qwen3_coder \
--load-format instanttensor \
--attention-backend flashinfer \
--kv-cache-dtype fp8 \
--enable-prefix-caching \
--enable-chunked-prefill \
--enable-auto-tool-choice \
--trust-remote-code

```

```

HF_HUB_OFFLINE=1 tool-eval-bench \
--base-url http://127.0.0.1:4321 \
--short --perf \
--bench-args "--tokenizer ./models/Qwen3.5-122B-Hybrid-INT4FP8"

```

GPU 99.22GiB • Warm-up 1449ms • Quality 100/100 • Responsiveness 48/100 • Median turn 3.1s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
❑	d0	c1	2,006	40.0	1,245	4,263
		c2	1,718	59.7	2,357	6,249
		c4	1,827	80.5	4,359	9,892
❑	d4096	c1	2,421	39.9	2,725	5,748
		c2	2,241	62.3	5,456	9,242
		c4	2,175	75.7	10,955	16,419
❑	d8192	c1	2,255	38.7	4,728	7,853
		c2	2,228	60.4	9,144	13,014
		c4	2,127	59.9	18,482	24,253
Per-stream (single request, per concurrent client)						
❑	—	c1	2,421	40.0	1,245	—
		c2	2,241	62.3	—	—
		c4	2,175	80.5	—	—

### 3.3.8 vLLM.dev + Qwen3.5-122B-Hybrid-INT4FP8 + MTP-4

```

HYBRID_PATH="./models/Qwen3.5-122B-Hybrid-INT4FP8"
HYBRID_NAME=$(basename "$HYBRID_PATH")

```

```

VLLM_IMAGE="i-dgx-spark-gb10:vllm0.19.1.dev"

docker run \
  --rm -it \
  --gpus all \
  --name ${HYBRID_NAME} \
  -v $(pwd)/${HYBRID_PATH}:/models/${HYBRID_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${HYBRID_NAME} \
    --served-model-name ${HYBRID_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --gpu-memory-utilization 0.8 \
    --max-model-len 262144 \
    --max-num-seqs 4 \
    --max-num-batched-tokens 16384 \
    --speculative-config '{"method": "mtp", "num_speculative_tokens": 4}' \
    --reasoning-parser qwen3 \
    --tool-call-parser qwen3_coder \
    --load-format instanttensor \
    --attention-backend flashinfer \
    --kv-cache-dtype fp8 \
    --enable-prefix-caching \
    --enable-chunked-prefill \
    --enable-auto-tool-choice \
    --trust-remote-code

HF_HUB_OFFLINE=1 tool-eval-bench \
  --base-url http://127.0.0.1:4321 \
  --short --perf \
  --bench-args "--tokenizer ./models/Qwen3.5-122B-Hybrid-INT4FP8"

```

GPU 99.27GiB • Warm-up 1455ms • Quality 97/100 • Responsiveness 47/100 • Median turn 3.3s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
❑	d0	c1	1,965	40.3	1,278	4,256
		c2	1,422	49.5	3,065	7,370
		c4	1,772	79.2	4,766	10,129
❑	d4096	c1	2,317	38.4	2,853	5,989
		c2	2,204	56.6	5,520	9,610
		c4	2,155	71.8	11,018	16,755
❑	d8192	c1	2,216	40.3	4,823	7,797
		c2	2,169	61.3	9,387	13,159
		c4	2,108	56.4	18,593	24,845
Per-stream (single request, per concurrent client)						
❑	—	c1	2,317	40.3	4,823	—
		c2	2,204	61.3	—	—
		c4	2,155	79.2	—	—

### 3.3.9 vLLM.dev + Qwen3.5-122B-Hybrid-INT4FP8 + MTP-8

```
HYBRID_PATH="./models/Qwen3.5-122B-Hybrid-INT4FP8"
HYBRID_NAME=$(basename "$HYBRID_PATH")

VLLM_IMAGE="i-dgx-spark-gb10:vllm0.19.1.dev"

docker run \
  --rm -it \
  --gpus all \
  --name ${HYBRID_NAME} \
  -v $(pwd)/${HYBRID_PATH}:/models/${HYBRID_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
  serve /models/${HYBRID_NAME} \
  --served-model-name ${HYBRID_NAME} \
  --host 0.0.0.0 \
  --port 4321 \
  --gpu-memory-utilization 0.8 \
```

```

--max-model-len 262144 \
--max-num-seqs 4 \
--max-num-batched-tokens 16384 \
--speculative-config '{"method": "mtp", "num_speculative_tokens": 8}' \
--reasoning-parser qwen3 \
--tool-call-parser qwen3_coder \
--load-format instanttensor \
--attention-backend flashinfer \
--kv-cache-dtype fp8 \
--enable-prefix-caching \
--enable-chunked-prefill \
--enable-auto-tool-choice \
--trust-remote-code

```

```

HF_HUB_OFFLINE=1 tool-eval-bench \
--base-url http://127.0.0.1:4321 \
--short --perf \
--bench-args "--tokenizer ./models/Qwen3.5-122B-Hybrid-INT4FP8"

```

GPU 99.44GiB • Warm-up 1580ms • Quality 73/100 • Responsiveness 44/100 • Median turn 3.5s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
■	d0	c1	1,875	31.8	1,361	5,149
		c2	1,297	48.6	3,708	8,616
		c4	2,051	77.7	3,911	9,633
■	d4096	c1	2,337	29.4	2,864	6,986
		c2	2,165	48.4	5,585	10,294
		c4	2,112	60.5	11,192	17,965
■	d8192	c1	2,181	28.6	4,931	9,177
		c2	2,152	47.2	9,423	14,196
		c4	2,093	50.8	18,673	25,235
Per-stream (single request, per concurrent client)						
■	—	c1	2,337	31.8	1,361	—
		c2	2,165	48.6	—	—
		c4	2,112	77.7	—	—

### 3.4 补充测试

#### 3.4.1 vLLM.dev + Qwen3.6-27B

```
MODEL_PATH="./models/Qwen3.6-27B"
```

```

MODEL_NAME=$(basename "$MODEL_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm0.19.1.dev"

docker run \
  --rm -it \
  --gpus all \
  --name ${MODEL_NAME} \
  -v ${MODEL_PATH}:/models/${MODEL_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${MODEL_NAME} \
    --served-model-name ${MODEL_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --gpu-memory-utilization 0.8 \
    --max-model-len 262144 \
    --max-num-seqs 4 \
    --max-num-batched-tokens 16384 \
    --reasoning-parser qwen3 \
    --tool-call-parser qwen3_coder \
    --load-format instanttensor \
    --attention-backend flashinfer \
    --kv-cache-dtype fp8 \
    --enable-prefix-caching \
    --enable-chunked-prefill \
    --enable-auto-tool-choice \
    --trust-remote-code

HF_HUB_OFFLINE=1 tool-eval-bench \
  --base-url http://127.0.0.1:4321 \
  --short --perf \
  --bench-args "--tokenizer ./models/Qwen3.6-27B"

```



GPU 96.49GiB • Warm-up 1282ms • Quality 93/100 • Responsiveness 4/100 • Median turn 26.5s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
🟦	d0	c1	1,339	4.4	1,841	30,399
		c2	1,209	8.6	3,264	32,563
		c4	1,311	16.7	6,095	35,952
🟪	d4096	c1	1,374	4.4	4,778	33,372
		c2	1,338	8.4	8,871	38,500
		c4	1,367	16.0	17,479	47,922
🟩	d8192	c1	1,358	4.4	7,846	36,498
		c2	1,331	8.5	15,188	44,846
		c4	1,348	13.5	28,343	60,318
Per-stream (single request, per concurrent client)						
🟨	—	c1	1,374	4.4	1,841	—
		c2	1,338	8.6	—	—
		c4	1,367	16.7	—	—

### 3.4.2 vLLM.dflash + Qwen3.6-27B-NVFP4 + DFlash-8

```

TARGET_PATH="./models/Qwen3.6-27B-NVFP4"
DRAFT_PATH="./models/Qwen3.6-27B-DFlash"
TARGET_NAME=$(basename "$TARGET_PATH")
DRAFT_NAME=$(basename "$DRAFT_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm.dflash"

docker run \
  --rm -it \
  --gpus all \
  --name ${TARGET_NAME}-dflash \
  -v $(pwd)/${TARGET_PATH}:/models/${TARGET_NAME} \
  -v $(pwd)/${DRAFT_PATH}:/models/${DRAFT_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${TARGET_NAME} \
    --served-model-name ${TARGET_NAME} \
    --host 0.0.0.0 \

```

```

--port 4321 \
--gpu-memory-utilization 0.8 \
--max-model-len 262144 \
--max-num-seqs 4 \
--max-num-batched-tokens 16384 \
--speculative-config '{"model": "/models/'${DRAFT_NAME}'", "method": "dflash",
"num_speculative_tokens": 8}' \
--reasoning-parser qwen3 \
--tool-call-parser qwen3_coder \
--attention-backend FLASH_ATTN \
--enable-prefix-caching \
--enable-chunked-prefill \
--enable-auto-tool-choice \
--trust-remote-code

```

```

HF_HUB_OFFLINE=1 tool-eval-bench \
--base-url http://127.0.0.1:4321 \
--short --perf \
--bench-yargs "--tokenizer ./models/Qwen3.6-27B-NVFP4"

```

GPU 96.19GiB • Warm-up 52326ms • Quality 97/100 • Responsiveness 37/100 • Median turn 4.3s						
Depth		Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
<i>Aggregate throughput (pp2048 / tg128)</i>						
▣	d0	c1	1,942	29.6	<b>1,337</b>	<b>5,425</b>
		c2	1,437	49.3	3,482	7,983
		c4	1,991	<b>87.5</b>	4,043	8,877
▣	d4096	c1	<b>2,037</b>	30.1	3,248	7,268
		c2	1,896	60.4	6,440	10,190
		c4	1,815	84.0	13,454	18,418
▣	d8192	c1	1,912	30.8	5,586	9,507
		c2	1,871	47.6	10,854	15,621
		c4	1,746	73.5	23,245	28,507
<i>Per-stream (single request, per concurrent client)</i>						
▣	—	c1	2,037	30.8	5,586	—
		c2	1,896	60.4	—	—
		c4	1,991	87.5	—	—

### 3.4.3 vLLM.dflash + Qwen3.6-27B-NVFP4 + DFlash-15

```

TARGET_PATH="./models/Qwen3.6-27B-NVFP4"
DRAFT_PATH="./models/Qwen3.6-27B-DFlash"

```

```

TARGET_NAME=$(basename "$TARGET_PATH")
DRAFT_NAME=$(basename "$DRAFT_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm.dflash"

docker run \
  --rm -it \
  --gpus all \
  --name ${TARGET_NAME}-dflash \
  -v $(pwd)/${TARGET_PATH}:/models/${TARGET_NAME} \
  -v $(pwd)/${DRAFT_PATH}:/models/${DRAFT_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${TARGET_NAME} \
    --served-model-name ${TARGET_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --gpu-memory-utilization 0.8 \
    --max-model-len 262144 \
    --max-num-seqs 4 \
    --max-num-batched-tokens 16384 \
    --speculative-config '{"model": "/models/'${DRAFT_NAME}', "method": "dflash",
    "num_speculative_tokens": 15}' \
    --reasoning-parser qwen3 \
    --tool-call-parser qwen3_coder \
    --attention-backend FLASH_ATTN \
    --enable-prefix-caching \
    --enable-chunked-prefill \
    --enable-auto-tool-choice \
    --trust-remote-code

HF_HUB_OFFLINE=1 tool-eval-bench \
  --base-url http://127.0.0.1:4321 \
  --short --perf \
  --bench-args "--tokenizer ./models/Qwen3.6-27B-NVFP4"

```

GPU 96.86GiB • Warm-up 46755ms • Quality 90/100 • Responsiveness 33/100 • Median turn 4.8s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
<i>Aggregate throughput (pp2048 / tg128)</i>						
□	d0	c1	1,492	33.0	<b>1,699</b>	<b>5,309</b>
		c2	1,734	49.3	2,289	6,756
		c4	<b>1,943</b>	69.0	4,118	9,603
□	d4096	c1	1,852	35.3	3,593	6,941
		c2	1,751	55.4	7,005	10,947
		c4	1,701	<b>71.0</b>	14,305	19,412
□	d8192	c1	1,938	26.2	5,558	10,171
		c2	1,873	42.8	10,845	15,536
		c4	1,750	63.9	23,097	28,586
<i>Per-stream (single request, per concurrent client)</i>						
□	—	c1	1,938	35.3	3,593	—
		c2	1,873	55.4	—	—
		c4	1,943	71.0	—	—

#### 3.4.4 vLLM.dflash + Qwen3.6-27B-AWQ-INT4 + DFlash-15

```

TARGET_PATH="./models/Qwen3.6-27B-AWQ-INT4"
DRAFT_PATH="./models/Qwen3.6-27B-DFlash"
TARGET_NAME=$(basename "$TARGET_PATH")
DRAFT_NAME=$(basename "$DRAFT_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm.dflash"

docker run \
  --rm -it \
  --gpus all \
  --name ${TARGET_NAME}-dflash \
  -v $(pwd)/${TARGET_PATH}:/models/${TARGET_NAME} \
  -v $(pwd)/${DRAFT_PATH}:/models/${DRAFT_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
  serve /models/${TARGET_NAME} \
  --served-model-name ${TARGET_NAME} \
  --host 0.0.0.0 \

```

```

--port 4321 \
--gpu-memory-utilization 0.8 \
--max-model-len 262144 \
--max-num-seqs 4 \
--max-num-batched-tokens 16384 \
--speculative-config '{"model": "/models/'${DRAFT_NAME}'", "method": "dflash",
"num_speculative_tokens": 15}' \
--reasoning-parser qwen3 \
--tool-call-parser qwen3_coder \
--attention-backend FLASH_ATTN \
--enable-prefix-caching \
--enable-chunked-prefill \
--enable-auto-tool-choice \
--trust-remote-code

```

```

HF_HUB_OFFLINE=1 tool-eval-bench \
--base-url http://127.0.0.1:4321 \
--short --perf \
--bench-args "--tokenizer ./models/Qwen3.6-27B-AWQ-INT4"

```

GPU 96.43GiB • Warm-up 53867ms • Quality 100/100 • Responsiveness 35/100 • Median turn 4.5s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
■	d0	c1	757	38.9	2,999	6,022
		c2	783	53.0	5,203	9,144
		c4	895	80.0	9,110	14,260
■	d4096	c1	924	42.2	6,913	9,684
		c2	909	54.5	13,441	17,638
		c4	905	90.9	26,930	31,410
■	d8192	c1	929	32.9	11,291	14,923
		c2	912	54.2	22,381	26,246
		c4	920	65.7	44,008	49,078
Per-stream (single request, per concurrent client)						
■	—	c1	929	42.2	6,913	—
		c2	912	54.5	—	—
		c4	920	90.9	—	—

### 3.4.5 vLLM.dflash + Qwen3.6-27B-GPTQ-INT4 + DFlash-15

```

TARGET_PATH="./models/Qwen3.6-27B-GPTQ-INT4"
DRAFT_PATH="./models/Qwen3.6-27B-DFlash"

```

```

TARGET_NAME=$(basename "$TARGET_PATH")
DRAFT_NAME=$(basename "$DRAFT_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm.dflash"

docker run \
  --rm -it \
  --gpus all \
  --name ${TARGET_NAME}-dflash \
  -v $(pwd)/${TARGET_PATH}:/models/${TARGET_NAME} \
  -v $(pwd)/${DRAFT_PATH}:/models/${DRAFT_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${TARGET_NAME} \
    --served-model-name ${TARGET_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --gpu-memory-utilization 0.8 \
    --max-model-len 262144 \
    --max-num-seqs 4 \
    --max-num-batched-tokens 16384 \
    --speculative-config '{"model": "/models/'${DRAFT_NAME}'", "method": "dflash",
"num_speculative_tokens": 15}' \
    --reasoning-parser qwen3 \
    --tool-call-parser qwen3_coder \
    --attention-backend FLASH_ATTN \
    --enable-prefix-caching \
    --enable-chunked-prefill \
    --enable-auto-tool-choice \
    --trust-remote-code

HF_HUB_OFFLINE=1 tool-eval-bench \
  --base-url http://127.0.0.1:4321 \
  --short --perf \
  --bench-args "--tokenizer ./models/Qwen3.6-27B-GPTQ-INT4"

```

GPU 96.57GiB • Warm-up 52788ms • Quality 97/100 • Responsiveness 39/100 • Median turn 4.0s						
Depth		Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
❑	d0	c1	939	26.4	2,401	7,035
		c2	875	52.5	4,653	8,547
		c4	972	84.3	8,366	12,854
❑	d4096	c1	1,005	36.2	6,323	9,655
		c2	984	62.7	12,445	15,944
		c4	1,000	86.6	24,440	28,837
❑	d8192	c1	1,006	32.9	10,385	14,069
		c2	996	48.1	20,486	24,218
		c4	994	75.0	40,761	45,777
Per-stream (single request, per concurrent client)						
❑	—	c1	1,006	36.2	6,323	—
		c2	996	62.7	—	—
		c4	1,000	86.6	—	—

### 3.4.6 vLLM.dflash + Qwen3.6-27B-FP8 + DFlash-15

```

TARGET_PATH="./models/Qwen3.6-27B-FP8"
DRAFT_PATH="./models/Qwen3.6-27B-DFlash"
TARGET_NAME=$(basename "$TARGET_PATH")
DRAFT_NAME=$(basename "$DRAFT_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm.dflash"

docker run \
  --rm -it \
  --gpus all \
  --name ${TARGET_NAME}-dflash \
  -v $(pwd)/${TARGET_PATH}:/models/${TARGET_NAME} \
  -v $(pwd)/${DRAFT_PATH}:/models/${DRAFT_NAME} \
  -p 4321:4321 \
  --ipc=host \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 10s \
  --health-timeout 5s \
  --health-retries 60 \
  --health-start-period 900s \
  --entrypoint vllm ${VLLM_IMAGE} \
    serve /models/${TARGET_NAME} \
    --served-model-name ${TARGET_NAME} \
    --host 0.0.0.0 \

```

```

--port 4321 \
--gpu-memory-utilization 0.8 \
--max-model-len 262144 \
--max-num-seqs 4 \
--max-num-batched-tokens 16384 \
--speculative-config '{"model": "/models/'${DRAFT_NAME}'", "method": "dflash",
"num_speculative_tokens": 15}' \
--reasoning-parser qwen3 \
--tool-call-parser qwen3_coder \
--attention-backend FLASH_ATTN \
--enable-prefix-caching \
--enable-chunked-prefill \
--enable-auto-tool-choice \
--trust-remote-code

```

```

HF_HUB_OFFLINE=1 tool-eval-bench \
--base-url http://127.0.0.1:4321 \
--short --perf \
--bench-args "--tokenizer ./models/Qwen3.6-27B-FP8"

```

GPU 97.02GiB • Warm-up 49748ms • Quality 100/100 • Responsiveness 27/100 • Median turn 5.7s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
■	d0	c1	1,746	23.0	1,439	6,742
		c2	1,014	46.8	4,348	9,130
		c4	636	63.7	12,927	19,434
■	d4096	c1	871	31.3	7,311	11,153
		c2	559	45.6	21,902	26,833
		c4	366	64.3	66,974	72,823
■	d8192	c1	918	29.6	11,408	15,478
		c2	429	45.0	47,600	52,530
		c4	342	49.0	118,998	125,535
Per-stream (single request, per concurrent client)						
■	—	c1	1,746	31.3	7,311	—
		c2	1,014	46.8	—	—
		c4	636	64.3	—	—

### 3.4.7 vLLM.dflash + Qwen3.6-27B-AEON-NVFP4-MTP-XS + DFlash-15

```

TARGET_PATH="./models/Qwen3.6-27B-AEON-Ultimate-Uncensored-Multimodal-NVFP4-MTP-XS"
DRAFT_PATH="./models/Qwen3.6-27B-DFlash"

```



```

TARGET_NAME=$(basename "$TARGET_PATH")
DRAFT_NAME=$(basename "$DRAFT_PATH")
VLLM_IMAGE="i-dgx-spark-gb10:vllm.dflash"

docker run \
  --rm -it \
  --gpus all \
  --ulimit memlock=-1 \
  --name ${TARGET_NAME}-DFlash \
  -p 4321:4321 \
  --ipc host \
  -v $(pwd)/${TARGET_PATH}:/models/${TARGET_NAME} \
  -v $(pwd)/${DRAFT_PATH}:/models/${DRAFT_NAME} \
  -e VLLM_ALLOW_LONG_MAX_MODEL_LEN=1 \
  -e TORCH_CUDA_ARCH_LIST=12.1a \
  -e PYTORCH_CUDA_ALLOC_CONF=expandable_segments:True \
  -e TORCH_MATMUL_PRECISION=high \
  -e NVIDIA_FORWARD_COMPAT=1 \
  -e NVIDIA_DISABLE_REQUIRE=1 \
  -e ENABLE_NVFP4_SM100=0 \
  -e VLLM_USE_FLASHINFER_MOE_FP4=0 \
  -e VLLM_TEST_FORCE_FP8_MARLIN=0 \
  -e VLLM_USE_FLASHINFER_SAMPLER=1 \
  -e VLLM_NVFP4_GEMM_BACKEND=flashinfer-cutlass \
  --health-cmd 'curl -sf http://localhost:4321/health || exit 1' \
  --health-interval 30s \
  --health-timeout 10s \
  --health-retries 3 \
  --health-start-period 600s \
  ${VLLM_IMAGE} vllm \
    serve /models/${TARGET_NAME} \
    --served-model-name ${TARGET_NAME} \
    --host 0.0.0.0 \
    --port 4321 \
    --tensor-parallel-size 1 \
    --dtype auto \
    --quantization modelopt \
    --kv-cache-dtype auto \
    --max-model-len 262144 \
    --max-num-seqs 64 \
    --max-num-batched-tokens 32768 \
    --gpu-memory-utilization 0.75 \
    --enable-chunked-prefill \
    --no-enable-prefix-caching \
    --generation-config vllm \
    --load-format safetensors \
    --trust-remote-code \
    --enable-auto-tool-choice \
    --tool-call-parser qwen3_coder \
    --reasoning-parser qwen3 \
    --attention-backend flash_attn \
    --limit-mm-per-prompt '{"image": 4, "video": 2}' \
    --mm-encoder-tp-mode data \
    --mm-processor-cache-type shm \

```

```

--mm-shm-cache-max-object-size-mb 256 \
--speculative-config '{"model": "/models/'${DRAFT_NAME}'"', "method": "dflash",
"num_speculative_tokens": 15}'

HF_HUB_OFFLINE=1 tool-eval-bench \
--base-url http://127.0.0.1:4321 \
--short --perf \
--bench-args "--tokenizer ./models/Qwen3.6-27B-AEON-Ultimate-Uncensored-Multimodal-NVFP4-MTP-XS"

```

GPU 92.45GiB • Warm-up 87219ms • Quality 83/100 • Responsiveness 13/100 • Median turn 10.7s

	Depth	Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
🟩	d0	c1	990	16.7	<b>2,832</b>	<b>10,078</b>
		c2	976	21.1	4,266	14,219
		c4	<b>1,165</b>	<b>32.1</b>	7,034	19,525
🟪	d4096	c1	741	17.3	9,774	16,756
		c2	956	16.1	13,212	27,208
		c4	973	27.3	25,737	39,939
🟨	d8192	c1	987	7.1	11,019	28,707
		c2	1,057	19.5	19,362	31,970
		c4	943	13.4	41,352	70,297
Per-stream (single request, per concurrent client)						
🟨	—	c1	1,249	17.3	9,774	—
		c2	1,072	21.1	—	—
		c4	1,165	32.1	—	—

### 3.4.8 LLaMA.cpp + Qwen3.6-27B-UD-Q4\_K\_XL + MTP-2

```

llama-server \
-m ./models/Qwen3.6-27B-UD-Q4_K_XL.gguf \
--alias Qwen3.6-27B-UD-Q4_K_XL.gguf \
--host 127.0.0.1 \
--port 4321 \
--ngl 99 \
-c 262144 \
-b 8192 \
-ub 4096 \
-np 1 \
-t -1 \
--cache-ram 0 \
--cache-reuse 512 \

```

```

-fa on \
--reasoning on \
--spec-type draft-mtp \
--spec-draft-n-max 2 \
--jinja

HF_HUB_OFFLINE=1 tool-eval-bench \
--base-url http://127.0.0.1:4321 \
--short --perf \
--bench-args "--tokenizer ./models/Qwen3.6-27B"

```

GPU 39.32GiB • Warm-up 1529ms • Quality 93/100 • Responsiveness 12/100 • Median turn 11.2s						
Depth		Concurrency	pp t/s	tg t/s	TTFT (ms)	Total (ms)
Aggregate throughput (pp2048 / tg128)						
▣	d0	c1	290	14.5	7,508	15,889
		c2	177	10.4	15,321	23,488
		c4	149	9.0	31,359	39,471
▣	d4096	c1	279	15.4	22,458	30,302
		c2	233	6.5	37,590	45,602
		c4	215	5.0	68,362	76,473
▣	d8192	c1	274	15.9	37,824	45,449
		c2	236	4.6	62,347	64,508
		c4	233	3.5	106,886	114,957
Per-stream (single request, per concurrent client)						
▣	—	c1	290	15.9	37,824	—
		c2	236	10.4	—	—
		c4	233	9.0	—	—

#### 3.4.9 Qwen3.6-27B-ParoQuant + NVFP4

#### 3.4.10 Qwen3.6-27B-TQ3\_4S + MTP

#### 3.4.11 vLLM.dev + Qwen3.5-122B-A10B-INT4-AutoRound-Intel + MTP-2 + Long-Text

#### 3.4.12 vLLM.dev + Qwen3.5-122B-Hybrid-INT4FP8 + MTP-2 + Long-Text