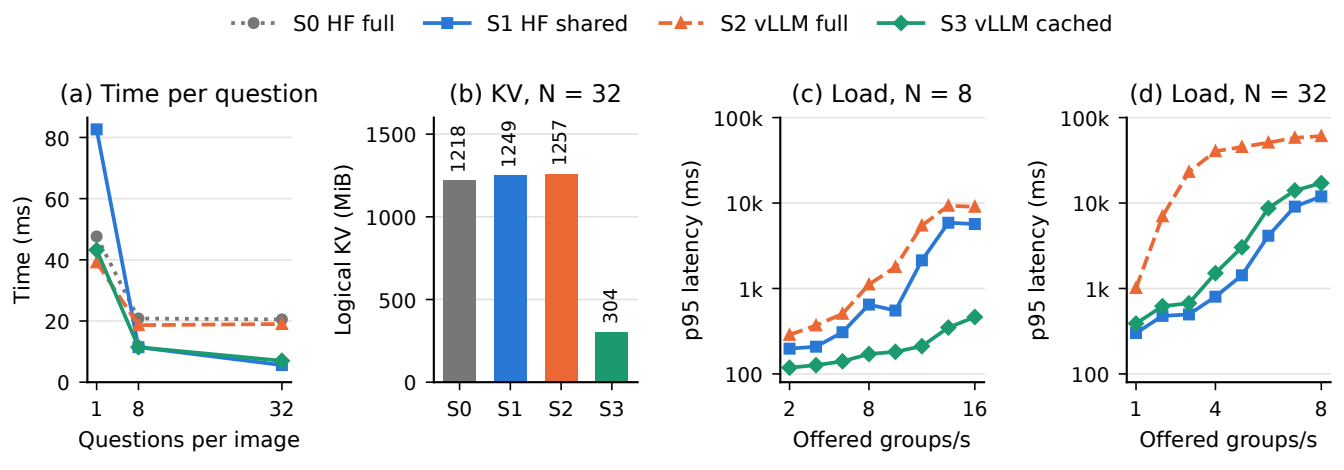

S0 HF full
S1 HF shared
S2 vLLM full
S3 vLLM cached
(a) Time per question
Time (ms)
Questions per image
(b) KV, N = 32
Logical KV (MiB)
1218
1249
1257
304
S0
S1
S2
S3
(c) Load, N = 8
p95 latency (ms)
Offered groups/s
(d) Load, N = 32
p95 latency (ms)
Offered groups/s