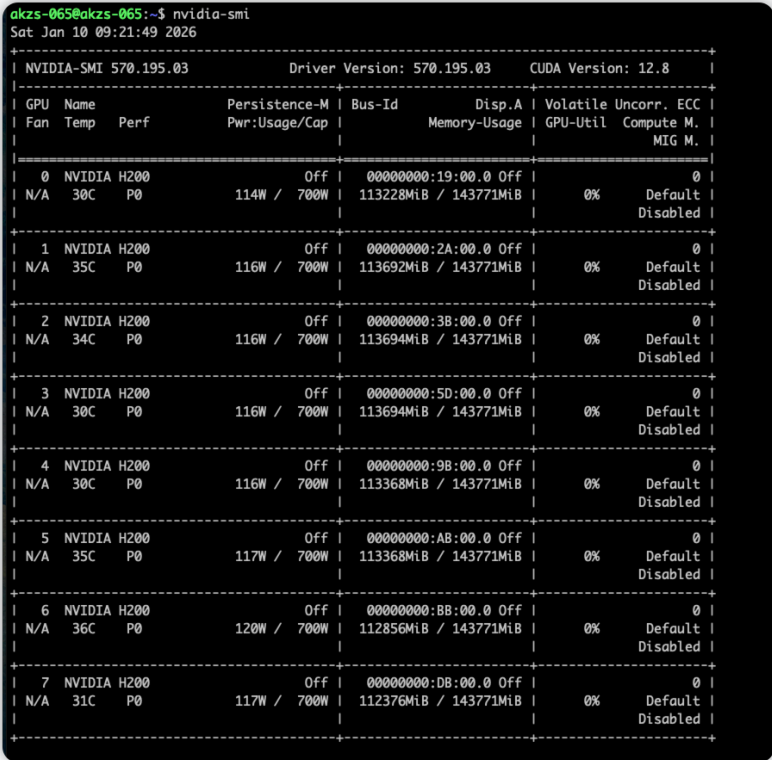


1. 基本信息

项目描述	测试趋境科技 8 卡 H200 方案 DeepSeek-V3.2 综合表现 (精度/性能)
测试时间	2026.01.10

1. 测试环境

1. 硬件配置

硬件名称	配置信息	数量	
CPU	INTEL(R) XEON(R) PLATINUM 8558	2	
GPU	 <pre>akzs-065@akzs-065:~\$ nvidia-smi Sat Jan 10 09:21:49 2026 +-----+ NVIDIA-SMI 570.195.03 Driver Version: 570.195.03 CUDA Version: 12.8 +-----+-----+ GPU Name Persistence-M Bus-Id Disp.A Volatile Uncorr. ECC Fan Temp Perf Pwr:Usage/Cap Memory-Usage GPU-Util Compute M. =====+=====+ 0 NVIDIA H200 Off 00000000:19:00.0 Off 0 N/A 30C P0 114W / 700W 113228MiB / 143771MiB 0% Default +-----+-----+ 1 NVIDIA H200 Off 00000000:2A:00.0 Off 0 N/A 35C P0 116W / 700W 113692MiB / 143771MiB 0% Default +-----+-----+ 2 NVIDIA H200 Off 00000000:3B:00.0 Off 0 N/A 34C P0 116W / 700W 113694MiB / 143771MiB 0% Default +-----+-----+ 3 NVIDIA H200 Off 00000000:5D:00.0 Off 0 N/A 30C P0 116W / 700W 113694MiB / 143771MiB 0% Default +-----+-----+ 4 NVIDIA H200 Off 00000000:9B:00.0 Off 0 N/A 30C P0 116W / 700W 113368MiB / 143771MiB 0% Default +-----+-----+ 5 NVIDIA H200 Off 00000000:AB:00.0 Off 0 N/A 35C P0 117W / 700W 113368MiB / 143771MiB 0% Default +-----+-----+ 6 NVIDIA H200 Off 00000000:BB:00.0 Off 0 N/A 36C P0 120W / 700W 112856MiB / 143771MiB 0% Default +-----+-----+ 7 NVIDIA H200 Off 00000000:DB:00.0 Off 0 N/A 31C P0 117W / 700W 112376MiB / 143771MiB 0% Default +-----+-----+</pre>	8	
内存	类型	DDR5	32
单条容量	64 GB		
上述机器数量	3		

1. 软件配置

软件名称	版本信息
操作系统	Ubuntu 22.04.5 LTS (Jammy Jellyfish)
内核版本	5.15.0-164-generic
测试工具	evalscope v1.4.1

2. 测试用例

1. 性能测试

测试目的	验证上述趋境方案 DeepSeek-V3.2 性能表现（TTFT/TPS等）符合预期。
测试过程	<pre>evalscope perf --rate 5 --parallel 84 --number 512 --model DeepSeek-V3.2 --tokenizer-path /data/models/DeepSeek-V3.2 --url http://127.0.0.1:8000/v1/completions --api openai --dataset random --max-tokens 1000 --min-tokens 1000 --prefix-length 0 --min-prompt-length 6000 --max-prompt-length 6000 --extra-args '{"ignore_eos": true}'</pre>
预期结果	TTFT ≤ 2s; TPOT ≤ 20ms

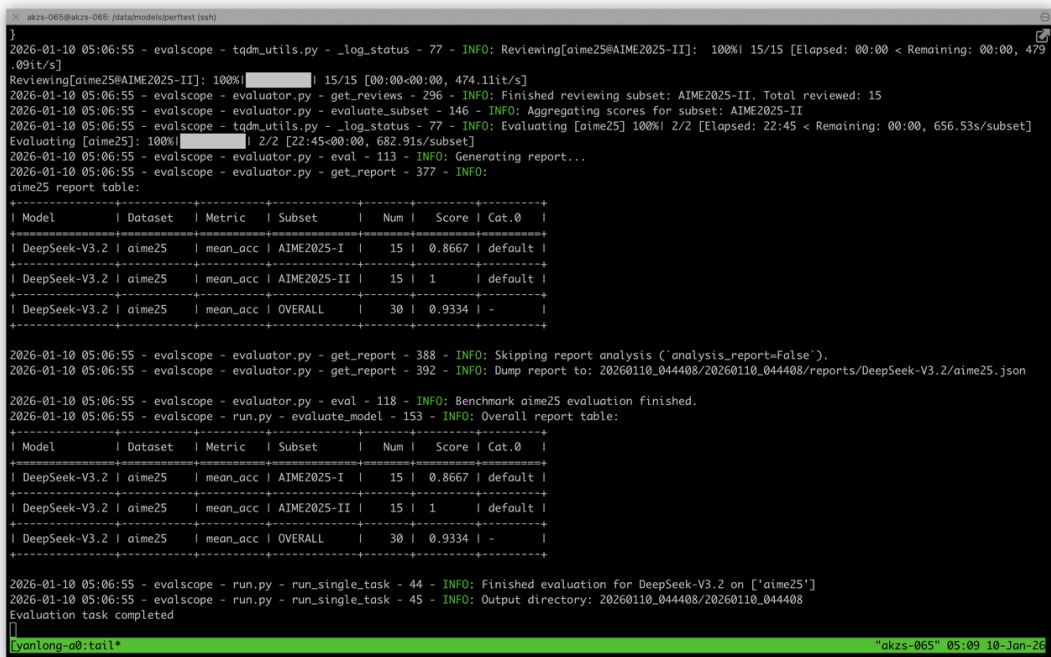
测试
结果

```
2026-01-10 09:33:15 - evalscope - INFO:
Benchmarking summary:
+-----+-----+
| Key | Value |
+-----+-----+
| Time taken for tests (s) | 157.897 |
+-----+-----+
| Number of concurrency | 84 |
+-----+-----+
| Request rate (req/s) | 5 |
+-----+-----+
| Total requests | 512 |
+-----+-----+
| Succeed requests | 512 |
+-----+-----+
| Failed requests | 0 |
+-----+-----+
| Output token throughput (tok/s) | 3320.44 |
+-----+-----+
| Total token throughput (tok/s) | 22776.2 |
+-----+-----+
| Request throughput (req/s) | 3.2426 |
+-----+-----+
| Average latency (s) | 22.3776 |
+-----+-----+
| Average time to first token (s) | 1.7905 |
+-----+-----+
| Average time per output token (s) | 0.0201 |
+-----+-----+
| Average inter-token latency (s) | 0.0638 |
+-----+-----+
| Average input tokens per request | 6000 |
+-----+-----+
| Average output tokens per request | 1024 |
+-----+-----+
2026-01-10 09:33:16 - evalscope - INFO:
Percentile results:
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| Percentiles | TTFT (s) | ITL (s) | TPOT (s) | Latency (s) | Input tokens | Output tokens | Output (tok/s) | Total (tok/s) |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 10% | 0.9769 | 0.0484 | 0.016 | 17.9088 | 6000 | 1024 | 35.9927 | 246.8875 |
| 25% | 1.2695 | 0.057 | 0.0167 | 18.9631 | 6000 | 1024 | 40.5589 | 278.2084 |
| 50% | 1.8004 | 0.0627 | 0.0191 | 21.353 | 6000 | 1024 | 48.0547 | 329.625 |
| 66% | 2.0093 | 0.0665 | 0.0214 | 23.5359 | 6000 | 1024 | 52.2707 | 358.5441 |
| 75% | 2.1204 | 0.0694 | 0.0228 | 25.2603 | 6000 | 1024 | 54.0283 | 370.6001 |
| 80% | 2.2 | 0.0717 | 0.0234 | 25.9075 | 6000 | 1024 | 54.9383 | 376.8422 |
| 90% | 2.5418 | 0.0776 | 0.0259 | 28.4502 | 6000 | 1024 | 57.1787 | 392.2101 |
| 95% | 2.9217 | 0.0925 | 0.0289 | 31.0918 | 6000 | 1024 | 58.6787 | 402.4989 |
| 98% | 3.6751 | 0.1246 | 0.0306 | 33.3687 | 6000 | 1024 | 62.6386 | 429.6615 |
| 99% | 4.3621 | 0.1342 | 0.0322 | 34.667 | 6000 | 1024 | 64.9052 | 445.2094 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
2026-01-10 09:33:16 - evalscope - INFO: Save the summary to: ./outputs/20260110_093001/appauto-bench-0-p84
Save the csv result to: 20260110_093001.csv
Save the xlsx result to: 20260110_093001.xlsx
```

结
论

通过

3. 正确性测试

测试目的	验证上述趋境方案 DeepSeek-V3.2 精度表现符合预期。
测试过程	<p>正确性测试 (aime25): evalscope eval --model DeepSeek-V3.2 --api-url http://127.0.0.1:8000/v1/ --api-key EMPTY --eval-type openai_api --datasets 'aime25' --stream --timeout 6000 --generation-config '{"do_sample":true,"temperature":0.6,"max_tokens":56000,"extra_body":{"chat_template_kwargs":{"thinking": true}}}' --dataset-args '{"aime25":{"prompt_template": "{question}\\nPlease reason step by step and place your final answer within boxed({}). Force Requirement: 1.only the final answer should be wrapped in boxed({}); 2.no other numbers or text should be enclosed in boxed({}). 3.Answer in English"}}' --eval-batch-size 16</p>
测试结果	 <pre>2026-01-10 05:06:55 - evalscope - tqdm_utils.py - _log_status - 77 - INFO: Reviewing[aime25@AIME2025-II]: 100% 15/15 [00:00<00:00, 474.11it/s] Reviewing[aime25@AIME2025-II]: 100% 15/15 [00:00<00:00, 474.11it/s] 2026-01-10 05:06:55 - evalscope - evaluator.py - get_reviews - 296 - INFO: Finished reviewing subset: AIME2025-II. Total reviewed: 15 2026-01-10 05:06:55 - evalscope - evaluator.py - evaluate_subset - 146 - INFO: Aggregating scores for subset: AIME2025-II 2026-01-10 05:06:55 - evalscope - tqdm_utils.py - _log_status - 77 - INFO: Evaluating [aime25] 100% 2/2 [22:45<00:00, 682.91s/subset] Evaluating [aime25]: 100% 2/2 [22:45<00:00, 682.91s/subset] 2026-01-10 05:06:55 - evalscope - evaluator.py - eval - 113 - INFO: Generating report... 2026-01-10 05:06:55 - evalscope - evaluator.py - get_report - 377 - INFO: aime25 report table: +-----+-----+-----+-----+-----+-----+-----+ Model Dataset Metric Subset Num Score Cat.0 +-----+-----+-----+-----+-----+-----+-----+ DeepSeek-V3.2 aime25 mean_acc AIME2025-I 15 0.8667 default +-----+-----+-----+-----+-----+-----+-----+ DeepSeek-V3.2 aime25 mean_acc AIME2025-II 15 1 default +-----+-----+-----+-----+-----+-----+-----+ DeepSeek-V3.2 aime25 mean_acc OVERALL 30 0.9334 - +-----+-----+-----+-----+-----+-----+-----+ 2026-01-10 05:06:55 - evalscope - evaluator.py - get_report - 388 - INFO: Skipping report analysis ('analysis_report=False'). 2026-01-10 05:06:55 - evalscope - evaluator.py - get_report - 392 - INFO: Dump report to: 20260110_044408/20260110_044408/reports/DeepSeek-V3.2/aime25.json 2026-01-10 05:06:55 - evalscope - evaluator.py - eval - 118 - INFO: Benchmark aime25 evaluation finished. 2026-01-10 05:06:55 - evalscope - run.py - evaluate_model - 153 - INFO: Overall report table: +-----+-----+-----+-----+-----+-----+-----+ Model Dataset Metric Subset Num Score Cat.0 +-----+-----+-----+-----+-----+-----+-----+ DeepSeek-V3.2 aime25 mean_acc AIME2025-I 15 0.8667 default +-----+-----+-----+-----+-----+-----+-----+ DeepSeek-V3.2 aime25 mean_acc AIME2025-II 15 1 default +-----+-----+-----+-----+-----+-----+-----+ DeepSeek-V3.2 aime25 mean_acc OVERALL 30 0.9334 - +-----+-----+-----+-----+-----+-----+-----+ 2026-01-10 05:06:55 - evalscope - run.py - run_single_task - 44 - INFO: Finished evaluation for DeepSeek-V3.2 on ['aime25'] 2026-01-10 05:06:55 - evalscope - run.py - run_single_task - 45 - INFO: Output directory: 20260110_044408/20260110_044408 Evaluation task completed [yanlong-a0:tail* "akzs-065" 05:09 10-Jan-26]</pre>
评分标准	官方分数
结论	通过 (趋境: 93.3; 官方: 93.1)

4. 超长上下文测试

测试目的	验证上述趋境方案 DeepSeek-V3.2 支持超长上下文。
测试过程	evalscope perf --parallel 1 --number 1 --model DeepSeek-V3.2 --tokenizer-path /data/models/DeepSeek-V3.2 --url http://127.0.0.1:8000/v1/completions --api openai --dataset random --max-tokens 1000 --min-tokens 1000 --prefix-length 0 --min-prompt-length 128000 --max-prompt-length 128000 --extra-args '{"ignore_eos": true}'
预期结果	支持超长上下文。

测试
结果

Benchmarking summary:

Key	Value
Time taken for tests (s)	38.7012
Number of concurrency	1
Request rate (req/s)	-1
Total requests	1
Succeed requests	1
Failed requests	0
Output token throughput (tok/s)	25.839
Total token throughput (tok/s)	3333.23
Request throughput (req/s)	0.0258
Average latency (s)	38.7012
Average time to first token (s)	24.2425
Average time per output token (s)	0.0145
Average inter-token latency (s)	0.0427
Average input tokens per request	128000
Average output tokens per request	1000

结
论

通过

1. 结论

- 三台服务器的TPM为 $(22776 * 60 = 1,366,560)$ ，TOPT，TTFT，精度均符合要求，结合上述测试结果，测试通过。当前性能仍在持续优化中。