PR #33944 完整报告

sgl-project/sglang

[CI] Share VLM engines and prune launch matrices on the per-commit H100/H200 suites

合并时间：2026-08-07 16:22

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/33944

执行摘要

- 一句话：共享 VLM engine，精简 H100/H200 CI 测试矩阵
- 推荐动作：值得负责 CI 与长测试套件的工程师精读。重点看三个设计决策：一是 _start_engine + tearDownClass 的组合如何用类级共享把 engine 启动成本摊薄，同时借 CustomTestCase 保证异常路径下的清理；二是 stage="extra-a" 的 label-gated 拆分方式与共享 fixture 常量的配套；三是“保持矩阵对角、skip 替代删除”的覆盖维护思路。普通功能开发者可略读，仅需知道 per-commit 测试结构有了变化。

功能与动机

PR body 明确说明了目标与边界："Trims per-commit wall time on the 1-gpu-h100, 4-gpu-h100 and 8-gpu-h200 pools by sharing server/engine launches, keeping matrix diagonals, and shrinking oversized eval sample counts. No accuracy threshold is changed."per-commit测试在这些主力池上耗时过长，VLM测试每个用例都单独启动Engine、Qwen3-Omni 等 30B 级模型服务启动代价很高，而部分 GSM8K 评估跑数千题只为一个二值结论。作者选择“共享启动 + 分层门控 + 压缩样本”来削减成本，而不是删覆盖或放宽准确率。

实现拆解

实现按 5 步拆解：

1. VLM 输入格式测试：每类共享一个 engine（核心改动）。test/registered/vlm/test_vlm_input_format.py 中把 Engine 创建从实例级 setUp 上移到类级 setUpClass 的新方法 _start_engine，清理从 tearDown 改为 tearDownClass；测试基类从 unittest.IsolatedAsyncioTestCase 换成 CustomTestCase，利用其“setUpClass 抛异常也执行 tearDownClass”的语义做引擎兜底回收。三个用例 test_accepts_image、test_accepts_precomputed_embeddings、test_accepts_processor_output 从 async_generate 改为同步 generate，原因是 tokenizer manager 会把 handle_loop 钉死在第一个请求所在的 loop 上，逐测试开 loop 会困死后续请求。est_time 从 747s 降到 300s，是本 PR 单项收益最大的改动。
2. 昂贵用例迁入 label-gated extra 阶段。新增 test/registered/sessions/test_streaming_session_swa_extra.py 与 test/registered/vlm/test_vision_openai_server_extra.py，均注册到 stage="extra-a"，只有打 run-ci-extra 标签才执行；per-commit 侧同步瘦身：test_streaming_session_swa.py 移除 TestStreamingSessionSWARetractMixedChunk（est_time 519→390），test_vision_openai_server_a.py 移除 TestQwen3OmniServer（

780→560）并为 --context-length 300 补校准注释。支撑改动是 python/sglang/test/server_fixtures/streaming_session_fixture.py 新增共享常量 SWA_MODEL、SWA_COMMON_ARGS，让 base 与 extra 两侧从同一处导入参数。

3. 压缩评估样本量。test/registered/models_e2e/test_minimax_m25_basic.py 的 GSM8K num_questions/parallel 从 1400 降到 200；test/registered/models_e2e/test_mimo_v2_flash.py 删除 gsm8k_num_questions = 1319 等覆盖改吃 kit 默认值（est_time 350→200）。所有准确率阈值原样保留，样本量下降约 7 倍会直接放大门控统计方差。
4. 保留矩阵对角，skip 替代删除。test/registered/models_e2e/test_qwen3_next_models_mtp.py 中 TestQwen3NextMTPV2 加 @unittest.skip 并标注 manual-only，覆盖映射为：topk=1 由 TestQwen3NextMTPLazyV2 覆盖、extra_buffer 策略由 TestQwen3NextMTPTopk 覆盖，该组合属对角线冗余。提交 d8c27367 曾尝试恢复 dsv32 tp/dp 测试，后因 #29686 要求删除对应文件而放弃。
5. 目录整理与 est_time 重标定。一批 server 端模型测试从 test/registered/models/ 迁到 test/registered/models_e2e/（test_inkling.py、test_generation_models.py、test_vlm_models.py 等），其中 test_inkling.py est_time 600→250；test_breakable_cuda_graph.py、test_post_capture_kv_sizing.py、test_lora_overlap_loading.py 等仅做 est_time 重标定。

关键文件：

- test/registered/vlm/test_vlm_input_format.py（模块 多模态输入；类别 test；类型 test-coverage；符号 setUp, _start_engine, tearDownClass, test_accepts_image）：核心改造：engine 从 per-test 改为 per-class 共享、async 改 sync、基类切换为 CustomTestCase，est_time 从 747s 降到 300s，是本 PR 收益最大的单项，也是后续共享 engine 模式的范本。
- test/registered/vlm/test_vision_openai_server_extra.py（模块 视觉服务；类别 test；类型 test-coverage；符号 TestQwen3OmniServer）：新增 label-gated extra 文件，承载被移出 per-commit 的 TestQwen3OmniServer，展示了 stage="extra-a" 的注册模式与 mixin 隔离写法。
- test/registered/sessions/test_streaming_session_swa_extra.py（模块 流式会话；类别 test；类型 test-coverage；符号 TestStreamingSessionSWARetractMixedChunk）：新增 label-gated extra 文件，SWA mixed-chunk retract 变体从 per-commit 迁入，与 base 文件共享 fixture 常量。
- python/sglang/test/server_fixtures/streaming_session_fixture.py（模块 会话夹具；类别 test；类型 test-coverage；符号 SWA_MODEL, SWA_COMMON_ARGS）：SWA_MODEL 与 SWA_COMMON_ARGS 上提为共享常量，供 per-commit 与 extra 两侧导入，是测试文件拆分的配套支撑，避免参数重复定义漂移。
- test/registered/vlm/test_vision_openai_server_a.py（模块 视觉服务；类别 test；类型 test-coverage；符号 TestQwen2VLContextLengthServer）：删除 TestQwen3OmniServer、重标 est_time（780→560），并为 context-length 300 校准补注释；提交历史显示此处曾尝试换 llava-0.5b 又因 warmup 越界回退。
- test/registered/models_e2e/test_qwen3_next_models_mtp.py（模块 MTP 测试；类别 test；类型 test-coverage；符号 TestQwen3NextMTPV2）：用 @unittest.skip 而非删除来

保留矩阵对角条目 TestQwen3NextMTPV2，并写明覆盖映射与 manual-only 定位，est_time 430→290。

- test/registered/models_e2e/test_minimax_m25_basic.py（模块 MiniMax 测试；类别 test；类型 test-coverage；符号 test_a_gsm8k）：GSM8K 样本量从 1400 降到 200（parallel 同步下调），是阈值不变前提下统计风险最集中的样本削减。
- test/registered/sessions/test_streaming_session_swa.py（模块 流式会话；类别 test；类型 test-coverage；符号 TestStreamingSessionSWA, TestStreamingSessionSWARetractLargePage, TestStreamingSessionSWAAbortLeakRepro）：per-commit 文件瘦身：移除 mixed-chunk 变体、改从 fixture 导入共享常量，est_time 519→390。

关键符号：_start_engine, tearDownClass, setUpClass, test_accepts_image, test_accepts_precomputed_embeddings, test_accepts_processor_output, TestStreamingSessionSWARetractMixedChunk, TestQwen3OmniServer

关键源码片段

test/registered/vlm/test_vlm_input_format.py

核心改造：engine 从 per-test 改为 per-class 共享、async 改 sync、基类切换为 CustomTestCase，est_time 从 747s 降到 300s，是本 PR 收益最大的单项，也是后续共享 engine 模式的范本。

```
class VLMInputTestBase:
    model_path = None
    chat_template = None
    processor = None
    visual = None # 可调用对象，用于预计算 embedding
    engine = None # 类级共享：每个测试类只启动一个 engine

    @classmethod
    def setUpClass(cls):
        assert cls.model_path is not None, "Set model_path in subclass"
        assert cls.chat_template is not None, "Set chat_template in subclass"

        cls.image_urls = [IMAGE_MAN_IRONING_URL, IMAGE_SGL_LOGO_URL]
        if _is_cuda:
            cls.device = torch.device("cuda")
        elif _is_xpu:
            cls.device = torch.device("xpu")
        else:
            cls.device = torch.device("cpu")

        cls.main_image = []
        for image_url in cls.image_urls:
            response = requests.get(image_url)
            cls.main_image.append(Image.open(BytesIO(response.content)))

        cls.processor = AutoProcessor.from_pretrained(
```

```
            cls.model_path, trust_remote_code=True, use_fast=True
        )
        _fix_added_tokens_encoding(cls.processor.tokenizer)
        cls._init_visual()
        cls._start_engine() # engine 启动从每个测试一次改为每类一次

    @classmethod
    def _start_engine(cls):
        # 一个类只启动一个 engine，所有测试顺序读它。这里刻意用同步
        # generate 而非 async：tokenizer manager 会把 handle_loop 钉死在
        # 第一个请求所在的 loop 上，若每个测试各自开 loop 就会困死。
        cls.engine = Engine(
            model_path=cls.model_path,
            chat_template=cls.chat_template,
            device=cls.device.type,
            mem_fraction_static=0.8,
            enable_multimodal=True,
            disable_cuda_graph=True,
            trust_remote_code=True,
        )

    @classmethod
    def tearDownClass(cls):
        # CustomTestCase 保证即使 setUpClass 抛异常也会走 tearDownClass，
        # 避免 engine 泄漏在 CI 机器上长期驻留。
        if cls.engine is not None:
            cls.engine.shutdown()
            cls.engine = None

    def test_accepts_image(self):
        req = self.get_completion_request()
        conv = generate_chat_conv(req, template_name=self.chat_template)
        text = conv.get_prompt()
        output = self.engine.generate(
            prompt=text,
            image_data=self.main_image,
            sampling_params=dict(temperature=0.0, max_new_tokens=512),
        )
        self.verify_response(output)
```

test/registered/vlm/test_vision_openai_server_extra.py

新增 label-gated extra 文件，承载被移出 per-commit 的 TestQwen3OmniServer，展示了 stage="extra-a" 的注册模式与 mixin 隔离写法。

```
"""Label-gated vision/omni server launches too expensive for the per-commit
budget; the per-commit set lives in test_vision_openai_server_a.py."""

import unittest
```

```
from sglang.test.ci.ci_register import register_cuda_ci
from sglang.test.vlm_utils import OmniOpenAITestMixin

# 注册到 extra-a 阶段：只有显式打 run-ci-extra 标签的 PR 才执行，
# 不再占用 per-commit 预算。
register_cuda_ci(est_time=180, stage="extra-a", runner_config="1-gpu-large")


class TestQwen3OmniServer(OmniOpenAITestMixin):
    model = "Qwen/Qwen3-Omni-30B-A3B-Instruct"
    extra_args = [ # H100 显存紧张的 workaround
        "--mem-fraction-static=0.90",
        "--disable-cuda-graph",
        "--disable-fast-image-processor",
        "--grammar-backend=none",
    ]


# 删除 mixin，避免它被当作独立测试类被收集执行。
del OmniOpenAITestMixin


if __name__ == "__main__":
    unittest.main()
```

python/sglang/test/server_fixtures/streaming_session_fixture.py

SWA_MODEL 与 SWA_COMMON_ARGS 上提为共享常量，供 per-commit 与 extra 两侧导入，是测试文件拆分的配套支撑，避免参数重复定义漂移。

```
# 共享的 SWA 测试常量：per-commit 与 extra 两个 streaming-session 测试文件
# 都从这里导入，避免相同参数在多个文件里重复定义导致漂移。
SWA_MODEL = "openai/gpt-oss-20b"

# gpt-oss-20b 的通用启动参数，与 TestSessionLatency/TestSWARadixCacheKL 对齐。
SWA_COMMON_ARGS = [
    "--mem-fraction-static",
    "0.70",
    "--cuda-graph-backend-prefill=disabled",
]

ABORT_REPRO_CONTEXT_LEN = 512
ABORT_REPRO_PAGE_SIZE = 256
ABORT_REPRO_GEN_LEN = 4
```

评论区精华

本 PR 没有任何 reviewer 评论或 review thread，作者自审自合并；真正的“评审”发生在 14 个 commit 与 bot 重跑里。值得注意的设计取舍：

- 提交 4e905f67 "revert context-length launch to qwen2-vl; llava-0.5b overruns warmup"：上下文长度用例曾想换更便宜的 llava-0.5b 压时间，实测 warmup 越界后回退到 qwen2-vl，并在代码里留下“必须高于 warmup 图展开长度、低于测试图展开长度”的校准注释。
- 提交 b38998f "restore the qwen3-next lazy extra buffer matrix; #31991 enabled the alloc-fail legs on purpose"：作者明确 #31991 是有意启用 alloc-fail 测试腿，裁剪矩阵时恢复并保留其意图。
- 提交 81d967e "keep MTPV2 and GLM-4.1V as skipped classes instead of deleting them"：skip 而非删除，保留本地手动回归能力。
- CI bot 的重跑名单覆盖了本 PR 改动的全部 11 个测试文件（含 vlm input format、vision openai server a/extra、streaming session swa/extra、breakable cuda graph、post capture kv sizing、inkling、qwen3 next mtp、mimo v2 flash、minimax m25 basic），全部通过。
- 无实质 review 讨论，CI 重跑覆盖全部改动文件 (other): 8 个 1-gpu-h100 与 1 个 4-gpu-h100 重跑全部通过，确认裁剪后的测试行为稳定；设计权衡主要由 14 个 commit 承载而非评审对话。

风险与影响

- 风险：技术风险集中在以下几点：
- 测试隔离性下降：test_vlm_input_format.py 改为类级共享 engine 后，若某个用例污染 engine 状态（异常中断、残留状态），会级联影响同类后续用例；目前依赖“所有测试只读”的假设，缺少显式状态校验。
- 样本量缩水的统计风险：test_minimax_m25_basic.py 的 GSM8K 从 1400 题降到 200 题而阈值不变，二项分布置信区间显著变宽，稳定过阈的模型可能偶发抖动失败，边缘回归被掩盖的概率上升；TestQwen3NextMTPV2 从 CI 移除后该组合只能靠相邻两条腿间接覆盖。
- extra 门控的覆盖依赖：Qwen3-Omni 服务与 SWA mixed-chunk retract 只跑 run-ci-extra，若流程漏打该标签，这些配置的回归会延迟到手动验证才暴露。
- 同步化依赖实现细节：async→sync 的正确性建立在“tokenizer manager 把 handle_loop 钉死在首个请求 loop”这一实现行为上，未来若重构事件循环，测试可能集体困死。
- 目录迁移的引用失效风险：多个文件从 test/registered/models/ 迁到 models_e2e/，若外部脚本或 CI 配置仍引用旧路径会静默漏跑。
- 影响：影响范围：
- 对贡献者：1-gpu-h100、4-gpu-h100、8-gpu-h200 三个 per-commit 主池墙钟时间显著下降，反馈循环加快，是全仓库贡献者都能感知的改进。
- 对覆盖策略：形成“per-commit 轻覆盖 + 对角线”与“extra 重覆盖”的两层结构，回归检测质量开始依赖 run-ci-extra 标签纪律。
- 对团队规范：引入了“共享 engine 需配合同步 generate”“skip 而非删除”“共享常量入 fixture”等 CI 测试编写约定，后续新增长测试应遵循同一模式。
- 无产品运行时影响：所有改动均为测试代码与 CI 注册信息，不涉及模型推理路径。

- 风险标记：共享 engine 降低测试隔离性，GSM8K 样本量缩减约 7 倍，per-commit 覆盖依赖 extra 标签纪律，async 转 sync 依赖事件循环钉死假设，测试目录迁移可能使旧路径引用失效

关联脉络

- PR #33929 [misc] Remove break-graph debug log; reclaim pid-less /dev/shm leaks in CI: 同属 per-commit CI 稳定与开销治理：清理无主 /dev/shm 泄漏与删除调试日志，与本次墙钟裁剪共同改善 CI 超时与资源雪崩问题。
- PR #33904 [CI] Refresh the CPU HF cache base only on main-ref runs: 同属 CI 预算优化线（缓存刷新条件收紧），与本次启动矩阵裁剪是同一治理方向的相邻改动。