PR #24623 完整报告

sgl-project/sglang

[test/fix]: isolate VLM MMMU eval output dirs to fix nightly-4-gpu cross-test pollution

合并时间：2026-05-09 06:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24623

执行摘要

- 一句话：隔离 VLM MMMU 输出目录，消除 nightly 跨测试污染
- 推荐动作：此 PR 是测试隔离和代码复用的优秀范例，值得推荐给所有参与测试维护的开发人员。特别值得关注的设计决策：使用 tempfile.TemporaryDirectory 替代固定路径，从源头消除并发 / 顺序污染；通过 Mixin 和基类抽取重复逻辑，减少了后续添加类似测试的重复工作。建议阅读 mmmu_vlm_kit.py 中的 MMMUMixin 和 MMMUMultiModelTestBase 实现。

功能与动机

自 2026-04-25 以来的每次 nightly-test-general-4-gpu-h100 运行中，test_encoder_dp.py::test_vlm_mmmu_benchmark 均以精度 0.4200 失败——这是读取陈旧文件的明显信号。根本原因是 test_encoder_dp.py 和 test_epd_disaggregation.py 都将 lmms-eval 结果写入共享的 ./logs/，而 PR #23518 将 EPD MMMU 测试移入同一 nightly-4-gpu 套件后，EPD 运行后残留的 JSON（精度 0.42）被 encoder_dp 的 glob.glob("./logs/**/*.json", recursive=True)[0] 意外拾取。CI 运行器从 Docker 切换到裸机后，./logs/ 不再被自动清理，使潜在 bug 暴露。

实现拆解

变更步骤如下：

1. 共享基类与 Mixin 优化：在 python/sglang/test/kits/mmmu_vlm_kit.py 中，为 MMMUMixin（单模型测试）和 MMMUMultiModelTestBase（多模型测试）统一结果查找逻辑，移除永不触发的非递归 fallback glob（recursive=True 已覆盖顶层文件）。
2. 重构 test_encoder_dp.py：删除大量内联代码（run_mmmu_eval、_run_vlm_mmmu_test、_read_output_from_files、setUpClass 等），直接继承 MMMUMultiModelTestBase，通过 other_args 定义 encoder_dp 特定参数。核心测试方法 test_vlm_mmmu_benchmark 使用 tempfile.TemporaryDirectory 确保每次运行都写入独立目录。
3. 重构 test_epd_disaggregation.py：三个 MMMU 测试类（TestEPDDisaggregationOneEncoder、TestEPDDisaggregationMultiEncoders、TestEPDDisaggregationGrpcEncoderMMMU）均改为继承 MMMUMixin，移除各自内联的 run_mmmu_eval 和 test_mmmu 方法，改用 mixin 提供的 test_mmmu。同样使用 tempfile.TemporaryDirectory 隔离输出路径。同时将 glob、json、tempfile 导入移到模块

顶部，符合 PEP 8。

4. 简化注释：在最终提交中剔除仅复述代码的冗余注释，提高可读性。

测试配套：变更仅涉及测试文件，未修改生产代码或配置。新增 tempfile 导入，移除不再需要的 os、glob、json 等导入。

关键文件：

- test/registered/vlm/test_encoder_dp.py（模块 编码器 DP；类别 test；类型 test-coverage；符号 TestVLMEncoderDP, setUpClass, run_mmmu_eval, _run_vlm_mmmu_test）：核心测试文件，从 CustomTestCase 迁移到 MMMUMultiModelTestBase，删除 227 行内联代码，使用临时目录隔离输出，是本次修复的主要受益者。
- test/registered/distributed/test_epd_disaggregation.py（模块 EPD 解耦；类别 test；类型 test-coverage；符号 TestEPDDisaggregationOneEncoder, run_mmmu_eval, test_mmmu, TestEPDDisaggregationMultiEncoders）：三个 EPD MMMU 测试类从内联实现迁移到 MMMUMixin，移除重复的 run_mmmu_eval 和 test_mmmu 方法，统一使用临时目录隔离输出；导入移至模块顶部。
- python/sglang/test/kits/mmmu_vlm_kit.py（模块 MMMU 测试工具；类别 test；类型 test-coverage）：共享测试工具模块，优化了结果文件查找逻辑，移除了冗余的 fallback glob，简化了 MMMUMixin 和 MMMUMultiModelTestBase 中的实现。

关键符号：TestVLMEncoderDP.test_vlm_mmmu_benchmark, MMMUMixin.test_mmmu, MMMUMultiModelTestBase._run_vlm_mmmu_test, TestEPDDisaggregationOneEncoder.test_mmmu, TestEPDDisaggregationMultiEncoders.test_mmmu, TestEPDDisaggregationGrpcEncoderMMMU.test_mmmu

关键源码片段

test/registered/vlm/test_encoder_dp.py

核心测试文件，从 CustomTestCase 迁移到 MMMUMultiModelTestBase，删除 227 行内联代码，使用临时目录隔离输出，是本次修复的主要受益者。

```
import random
import tempfile
import unittest
from types import SimpleNamespace

from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci
from sglang.test.kits.mmmu_vlm_kit import MMMUMultiModelTestBase
from sglang.test.test_utils import is_in_ci

register_cuda_ci(est_time=500, suite="nightly-4-gpu", nightly=True)
register_amd_ci(est_time=500, suite="nightly-amd-4-gpu", nightly=True)

MODELS = [
    SimpleNamespace(model="Qwen/Qwen2.5-VL-72B-Instruct", mmmu_accuracy=0.55),
```

```
    SimpleNamespace(model="Qwen/Qwen3-VL-32B-Instruct", mmmu_accuracy=0.55),
    SimpleNamespace(model="OpenGVLab/InternVL2_5-8B", mmmu_accuracy=0.52),
    SimpleNamespace(model="zai-org/GLM-4.1V-9B-Thinking", mmmu_accuracy=0.68),
]


class TestVLMEncoderDP(MMMUMultiModelTestBase):
    # --tp=4 覆盖基类默认值，--cuda-graph-max-bs 32 覆盖基类的 64
    other_args = [
        "--mm-enable-dp-encoder",
        "--tp=4",
        "--cuda-graph-max-bs",
        "32",
    ]

    def test_vlm_mmmu_benchmark(self):
        """运行所有模型，CI 中随机选一个以缩短时间"""
        models_to_test = MODELS
        if is_in_ci():
            models_to_test = [random.choice(MODELS)]
        for model in models_to_test:
            # 每个模型使用独立的临时目录，防止 `./logs/` 残留污染
            with tempfile.TemporaryDirectory(
                prefix=f"encoder_dp_{model.model.replace('/', '_')}_"
            ) as output_path:
                self._run_vlm_mmmu_test(model, output_path)


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

test/registered/distributed/test_epd_disaggregation.py

三个 EPD MMMU 测试类从内联实现迁移到 MMMUMixin，移除重复的 run_mmmu_eval 和 test_mmmu 方法，统一使用临时目录隔离输出；导入移至模块顶部。

```
@unittest.skipIf(is_in_ci(), "Skipping in CI to reduce multi-GPU runtime")
class TestEPDDisaggregationOneEncoder(MMMUMixin, PDDisaggregationServerBase):
    """Test EPD disaggregation with single encode server"""

    # Qwen2.5-VL-3B-Instruct scores ~0.40 on the 50-sample MMMU subset.
    accuracy = 0.40
    mmmu_args = ["--limit", "50"]

    @classmethod
    def setUpClass(cls):
        super().setUpClass()
        cls.model = DEFAULT_SMALL_VLM_MODEL_NAME_FOR_TEST
        cls.base_url = cls.lb_url # MMMUMixin 通过此变量构造 OPENAI_API_BASE
        cls.encode_port = f"{int(cls.lb_port) + 300}"
```

```
        cls.encode_url = f"http://{cls.base_host}:{cls.encode_port}"
        # ... 其余 setup 代码保持不变

    # test_mmmu 由 MMMUMixin 提供，使用 tempfile.TemporaryDirectory 隔离输出
    # 无需在每个测试类中重复实现
```

python/sglang/test/kits/mmmu_vlm_kit.py

共享测试工具模块，优化了结果文件查找逻辑，移除了冗余的 fallback glob，简化了 MMMUMixin 和 MMMUMultiModelTestBase 中的实现。

```
    def test_mmmu(self: CustomTestCase):
        """Run MMMU evaluation test."""
        with tempfile.TemporaryDirectory() as output_path:
            self.run_mmmu_eval(self.model, output_path)

            # recursive=True 足以匹配顶层文件和子目录，无需单独的 `*.json` fallback
            result_files = glob.glob(f"{output_path}/**/*.json", recursive=True)

            if not result_files:
                raise FileNotFoundError(f"No JSON result files found in {output_path}")

            result_file_path = result_files[0]
            with open(result_file_path, "r") as f:
                result = json.load(f)
                # ... 读取并断言精度
```

评论区精华

Reviewer gemini-code-assist[bot] 提出了 7 条评论，主要集中在：

- PEP 8 导入位置：建议将 glob、json 和 tempfile 移到文件顶部，而不是留在方法内。
- 冗余 fallback glob：指出 glob.glob(f"{output_path}/*.json") 在 recursive=True 下是多余的，因为 **/*.json 已覆盖顶层文件。
  - 作者在后续提交 #18978db 和 #b9cba44 中采纳了所有建议，包括移除冗余 glob、提升导入、精简注释。
  - PEP 8 导入位置与冗余 fallback glob (style): 作者采纳建议，在后续提交中将所有相关导入移到模块顶部，并删除了所有四处的 fallback glob。

风险与影响

- 风险：风险较低。变更仅限于测试文件，未触及任何运行时逻辑或模型代码。主要风险是重构后的测试可能在边缘场景中无法正确捕获失败，例如若 tempfile.TemporaryDirectory 在 CI 环境中不可写（但不可能），或 MMMUMixin/MMMUMultiModelTestBase 中的假设与特定测试类冲突（但已在多个模型中验证）。由于删除了大量重复代码，回归风险低。此外，移除非递归 fallback 可能在不常见的情况下（若 lmms-eval 升级改变输出结构）导致找不到结果文件，但 **/*.json 递归搜索覆盖绝大多数情况。

- 影响：用户 / 系统：无影响，未变更用户可见行为或 API。测试团队：显著提升 nightly 套件可靠性，消除因交叉污染导致的假阳性失败。开发人员：测试代码量大幅减少（净删除 433 行），代码复用性提高，更易于维护。影响范围仅限于两个测试文件和一个测试工具模块。
- 风险标记：测试隔离重构，非核心路径变更，中量代码删除

关联脉络

- PR #23518 Move EPD MMMU tests to nightly-4-gpu suite: 该 PR 将 EPD MMMU 测试移入 nightly-4-gpu 套件，导致与 *encoder_dp* 测试共享 ./logs/ 目录，从而暴露出交叉污染 bug。本 PR 正是为修复此问题而提出。