PR #7127 完整报告

PaddlePaddle/FastDeploy

[Others]add unit test

合并时间：2026-04-01 18:36

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7127

执行摘要

- 一句话：恢复并新增 V1 版本缓存管理和资源调度的单元测试文件。
- 推荐动作：建议工程师精读这些测试以理解 V1 调度和缓存机制，同时注意 review 中指出的配置不完整和资源清理问题，避免在类似测试中重复错误，并考虑删除伪造覆盖率的函数。

功能与动机

根据 PR body，动机是 ' 恢复误删文件 '。具体来说，是为了恢复并新增 V1 版本的单元测试，以避免功能回归并加强测试覆盖，确保缓存管理和资源调度模块的稳定性。

实现拆解

实现方案包括新增四个测试文件：1) tests/v1/cache_manager/test_encoder_cache.py 测试多模态编码器缓存的驱逐逻辑；2) tests/v1/cache_manager/test_prefix_cache.py 测试前缀缓存管理器的块匹配和多模态功能；3) tests/v1/test_resource_manager_v1.py 测试 ResourceManagerV1 的异步特征下载、抢占和扩展等路径；4) tests/v1/test_schedule_output.py 测试调度输出和输出缓存行为。

关键文件：

- tests/v1/cache_manager/test_encoder_cache.py（模块 KVCache）：测试 EncoderCacheManager，覆盖多模态缓存驱逐逻辑，是 KVCache 模块的关键测试。
- tests/v1/cache_manager/test_prefix_cache.py（模块 KVCache）：测试 PrefixCacheManager，验证块匹配和多模态前缀缓存，是 KVCache 模块的核心功能测试。
- tests/v1/test_resource_manager_v1.py（模块 Scheduler）：测试 ResourceManagerV1，覆盖异步特征下载、抢占和资源管理，是 Scheduler 模块的重要测试。
- tests/v1/test_schedule_output.py（模块 Scheduler）：测试调度输出和输出缓存行为，验证调度逻辑，是 Scheduler 模块的关键测试。

关键符号：test_mm_encoder_cache, test_normal_case, test_normal_schedule, test_force_coverage_lines

评论区精华

review 中，Copilot 指出测试配置不完整（如 model_cfg 缺少 architectures 和 version 字段）、资源未清理（共享内存和线程池未 shutdown）以及注释误导。fastdeploy-bot 标记了一个

严重问题：test_force_coverage_lines 函数伪造覆盖率数据，建议删除该函数。
EmmonsCurse 批准了 PR，表明问题可能被接受或后续处理。

- 测试配置不完整 (correctness): 建议补全 architectures 和 version 字段，或改用 EngineArgs.create_engine_config() 生成完整配置。
- 资源未清理 (testing): 建议在测试中添加清理代码，如使用 addCleanup 或 try-finally。
- 伪造覆盖率 (testing): 建议删除该函数，编写真实测试或使用 pragma 标注。

风险与影响

- 风险：技术风险包括：1) 测试代码配置不完整（如 tests/v1/test_schedule_output.py:36）可能导致初始化错误，影响测试可靠性；2) 资源未清理（如 tests/v1/test_schedule_output.py:89）可能导致共享内存残留和线程泄漏，干扰其他测试用例；3) 伪造覆盖率函数（tests/v1/test_resource_manager_v1.py:705）会掩盖真实覆盖缺口，降低 CI 指标可信度。
- 影响：对用户无直接影响；对系统，提高测试覆盖率有助于捕捉回归 bug，提升稳定性，但测试代码的问题可能引入额外维护负担；对团队，增加了测试维护工作，但通过 review 反馈优化了代码质量，促进最佳实践。
- 风险标记：测试资源泄漏，伪造覆盖率，配置不完整

关联脉络

- PR #7125 [Feature] Config eviction_duration: 涉及 KVCache 和 Scheduler 配置，本 PR 的测试文件可能覆盖相关缓存驱逐逻辑。
- PR #7107 [PD Disaggregation] Write the cache of preempted req to storage and refine PD Disaggregation: 优化 KVCache 和 Scheduler，与本 PR 的 ResourceManagerV1 和前缀缓存测试主题相关。
- PR #6929 [BugFix][KVCache] Fix mm hash boundary comparison in get_block_hash_extra_keys: 修复 KVCache bug，本 PR 包含 prefix cache 测试，可能验证类似功能。