

PR #31843 完整报告

sgl-project/sglang

[AMD] [CI] Enable 3 nested unit tests needing harness stub fixes

合并时间: 2026-08-10 16:09

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/31843>

执行摘要

- 一句话: AMD CI 新增 3 个嵌套单元测试, 修复 3 处测试 stub 缺口
- 推荐动作: 值得关注 `gpu_cases.py` 中 `_discover_unit_tests()` 的 HIP 分支设计, 这是 " 同一套测试在不同硬件车道差异化收集 " 的可复用方案: 对生产代码零侵入、对 CUDA 车道零影响。若只关心功能开发, 可快速浏览 PR body 的测试结果表; 若要扩展 AMD CI 覆盖, 建议精读本 PR 与 #31483 的配套改动。

功能与动机

在 #31483 中作者首次为 AMD 启用嵌套 `multimodal_gen` 单元测试, 并将需要代码修改才能通过的文件拆分为后续 PR。本 PR body 明确指出: "These stub gaps are pre-existing and would also fail on CUDA, so the fixes are portable rather than AMD-specific", 即这些 harness stub 缺陷并非 AMD 特有, 而是测试假对象与生产代码演进脱节所致; 修复具有可移植性, 但按既定策略仅在 AMD 车道启用。

实现拆解

1. 修复渐进分辨率测试的 stub 缺失: `python/sglang/multimodal_gen/test/unit/progressive_resolution/test_progressive.py` 中 3 个用例构造的 `ctx` 缺少 `cfg_policy` 字段, 而 `ProgressiveDenoisingStage._on_resolution_change()` 在 `cfg_policy` is None 时提前返回, 导致断言永远无法执行。修复为引入 `CFGPolicy` 类并在 3 处 `SimpleNamespace` 上补充 `cfg_policy=CFGPolicy()`。
2. 补齐 SanaWM 流式路径全局参数 stub: `python/sglang/multimodal_gen/test/unit/sana_wm/test_streaming_realtime_path.py` 的 `_global_args` fixture 缺少 `enable_breakable_cuda_graph` 字段, `DenoisingStage.__init__()` 会读取该属性。修复为在 `SimpleNamespace` 中增加 `enable_breakable_cuda_graph=False`。
3. 启用已修复的 Lingbot 因果去噪测试: `realtime/test_lingbot_causal_denoising.py` 的 stub 缺口 (`stage.transformer` 为 None 时 `_reset_causal_cache_config_defaults()` 应为 no-op) 已在上游修复, 本 PR 不修改该文件, 仅将其纳入扫描范围。
4. 注册 AMD 测试集合: `python/sglang/multimodal_gen/test/server/gpu_cases.py` 的 `_AMD_READY_NESTED_UNIT_TESTS` 追加 3 个文件路径; 该常量仅在 `_discover_unit_tests()` 的 `current_platform.is_hip()` 分支中被消费, CUDA 车道保持扁平 `test_*.py` glob 字节级不变。

5. 更新覆盖率归属: scripts/ci/utils/ci_coverage_report.py 的 `_MM_GEN_FILE_BACKENDS` 将 3 个文件映射为 ("AMD",), 确保 CI 覆盖率报告把这 3 个文件的覆盖结果记到 AMD 名下。
6. 验证: 在两个 ROCm 镜像 (7.0.0 / 7.2.0) 上运行定向 multimodal-gen-unit-test, 均 1118 passed、3 skipped; CUDA 对应作业通过, 确认无跨车道影响。

关键文件:

- python/sglang/multimodal_gen/test/unit/progressive_resolution/test_progressive.py (模块 渐进分辨率; 类别 test; 类型 test-coverage; 符号 `test_resolution_change_doubles_image_tokens`, `test_text_portion_is_unchanged_after_resolution_change`, `test_image_position_ids_use_grid_coordinates`): 3 个渐进分辨率单元测试的 stub 修复: 为 `_on_resolution_change` 的 ctx 注入非 None 的 `cfg_policy`, 解除因提前 return 导致的断言空转, 是本 PR 最有实质内容的一处测试改动。
- python/sglang/multimodal_gen/test/server/gpu_cases.py (模块 用例编排; 类别 test; 类型 test-coverage; 符号 `_AMD_READY_NESTED_UNIT_TESTS`, `_discover_unit_tests`): AMD 嵌套测试白名单 `_AMD_READY_NESTED_UNIT_TESTS` 追加 3 个文件, 并承载 `_discover_unit_tests()` 的 HIP-gated 收集逻辑, 是本 PR 在 CI 侧的核心开关。
- python/sglang/multimodal_gen/test/unit/sana_wm/test_streaming_realtime_path.py (模块 流式路径; 类别 test; 类型 test-coverage; 符号 `_global_args`): 为 `_global_args` fixture 补充 `enable_breakable_cuda_graph=False` 字段, 修复 `DenoisingStage.__init__` 读取该属性时的 `AttributeError`, 从而启用 4 个流式实时路径测试。
- scripts/ci/utils/ci_coverage_report.py (模块 覆盖率统计; 类别 infra; 类型 infrastructure; 符号 `_MM_GEN_FILE_BACKENDS`): 覆盖率分类器 `_MM_GEN_FILE_BACKENDS` 将 3 个文件映射为 AMD 归属, 确保 CI 报告中这 3 个文件的覆盖率被正确记到 AMD 名下而非 CUDA。

关键符号: `_discover_unit_tests`, `_global_args`, `test_resolution_change_doubles_image_tokens`, `test_text_portion_is_unchanged_after_resolution_change`, `test_image_position_ids_use_grid_coordinates`

关键源码片段

python/sglang/multimodal_gen/test/unit/progressive_resolution/test_progressive.py

3 个渐进分辨率单元测试的 stub 修复: 为 `_on_resolution_change` 的 ctx 注入非 None 的 `cfg_policy`, 解除因提前 return 导致的断言空转, 是本 PR 最有实质内容的一处测试改动。

```
# 渐进分辨率测试的 stub 修复: `_on_resolution_change` 在 cfg_policy 为 None 时提前返回,
# 测试假对象中该字段原本缺失, 导致用例永远走不到断言。这里补上非 None 的 CFGPolicy 实例。
from sglang.multimodal_gen.runtime.distributed.cfg_policy import CFGPolicy
```

```
def test_resolution_change_doubles_image_tokens(self):
    B = 2
    old_grid_h, old_grid_w = 4, 4
```

```

new_grid_h, new_grid_w = 8, 8
max_text_tokens = 10

stage = object.__new__(Ideogram4ProgressiveDenoisingStage)
server_args = SimpleNamespace(
    pipeline_config=SimpleNamespace(
        patch_size=self._PATCH, ae_scale_factor=self._AE
    )
)

old_num_img = old_grid_h * old_grid_w # 16
new_num_img = new_grid_h * new_grid_w # 64
new_h_pixel = new_grid_h * self._SCALE
new_w_pixel = new_grid_w * self._SCALE

# cfg_policy 必须非 None: `_on_resolution_change` 在 cfg_policy 为 None 时提前返回,
# 会绕过待测的核心逻辑; 此处的值本身不被进一步使用, 仅用于通过守卫分支。
ctx = SimpleNamespace(
    latents=torch.zeros(B, new_num_img, self.IN_C),
    extra=self._make_ctx_extra(B, old_num_img, max_text_tokens),
    cfg_policy=CFGPolicy(),
)
batch = SimpleNamespace(
    extra={
        "ideogram4": self._make_ideogram_extra(
            B, old_grid_h, old_grid_w, max_text_tokens
        )
    }
)

stage._on_resolution_change(ctx, batch, server_args, new_h_pixel, new_w_pixel)

data = batch.extra["ideogram4"]
self.assertEqual(data["num_image_tokens"], new_num_img)
# ... 后续断言: position_ids / segment_ids / indicator 的 shape 均按新分辨率更新

```

python/sglang/multimodal_gen/test/server/gpu_cases.py

AMD 嵌套测试白名单 `_AMD_READY_NESTED_UNIT_TESTS` 追加 3 个文件, 并承载 `_discover_unit_tests()` 的 HIP-gated 收集逻辑, 是本 PR 在 CI 侧的核心开关。

嵌套 unit/ 测试的 AMD 白名单: 仅在 HIP 平台消费, CUDA 车道保持扁平 glob 不变。
 # 前 9 个由 #31483 启用 (无代码改动); 本 PR 新增的 3 个带小型 stub 修复。

```

_AMD_READY_NESTED_UNIT_TESTS = (
    "realtime/test_causal_denoising.py",
    "realtime/test_output_materialization.py",
    "realtime/test_realtime_consistency_harness.py",
    "realtime/test_realtime_control_signals.py",
    "realtime/test_realtime_output_transport.py",
    "realtime/test_realtime_vae.py",

```

```

"sana_wm/test_streaming_cached.py",
"sana_wm/test_streaming_stage.py",
"sana_wm/test_streaming_vae.py",
# 本 PR 新增: 需要 test-harness stub 修复后才能通过
"progressive_resolution/test_progressive.py",
"sana_wm/test_streaming_realtime_path.py",
# stub gap 已在上游修复, 仅需启用
"realtime/test_lingbot_causal_denoising.py",
)

def _discover_unit_tests() -> list[str]:
    unit_dir = Path(__file__).resolve().parent.parent / "unit"
    if not unit_dir.is_dir():
        return []
    # 扁平 unit/ 测试在所有平台运行, 保持 CUDA `multimodal-gen-unit-test` 作业字节级一致
    flat = [f"../unit/{f.name}" for f in unit_dir.glob("test_*.py") if f.is_file()]
    if not current_platform.is_hip():
        return sorted(flat)
    # 仅 AMD/ROCm 额外收集经过验证的嵌套子目录测试
    nested = [
        f"../unit/{rel}"
        for rel in _AMD_READY_NESTED_UNIT_TESTS
        if (unit_dir / rel).is_file()
    ]
    return sorted(flat + nested)

```

评论区精华

本 PR 无 review comment 争议, BBuf 直接 APPROVED。主要讨论来自 issue 评论:
[amd-bot](#) 指出标准 **PR Test (AMD)** 流程中 **multimodal-gen-unit-test-amd** 作业因 call-gate 取消而被跳过, 作者通过两次 **workflow_dispatch** 定向运行 (job 92704086756、92704227179) 完成验证, 并提醒审查者不能仅看主检查列表的 skip 状态。作者回应了 [amd-bot](#) 的 CI 状态说明并补充了 rebase 后的验证矩阵。

- 标准 PR Test 流程中覆盖作业被跳过 (testing): 变更本身验证充分 (两个 ROCm 车道均 1118 passed), 但标准管线的作业跳过会造成审查困惑; 作者补充了 run 链接作为证据。
- Rebase 收敛与验证矩阵 (other): 无争议; 验证记录完整, 覆盖了 AMD 双镜像与 CUDA 对照。

风险与影响

- 风险:
 1. 测试收集逻辑风险: `_discover_unit_tests()` 的 HIP 分支依赖 `current_platform.is_hip()` 判断, 若平台探测与实际运行环境不一致, 嵌套文件可能在错误车道被收集或遗漏; 当前 AMD 车道已有 12 个嵌套文件稳定运行, 风险较低。

2. 依赖上游修复: `realtime/test_lingbot_causal_denoising.py` 的启用依赖上游已合并的 `stage.transformer = None` stub 修复, 若上游回退或合入顺序错乱, 该文件会以未修复状态进入扫描范围而失败。
3. 继承的 main 失败: 两个 AMD 车道上 `test_output_materialization.py::test_file_path_transport_clears_in_memory_outputs` 已失败 (`GPUWorker.__new__` stub 缺少 #33725 引入的 `is_output_rank`), 该失败继承自 main、由 #31483 启用、本 PR 不触碰, 但会持续影响 AMD CI 绿灯状态。
4. 覆盖率归属准确性: `ci_coverage_report.py` 的路径与 `_discover_unit_tests()` 的相对路径若不一致, 会出现归属错记; 作者已通过真实报告验证 3 个文件正确归入 AMD。- 影响: 对 AMD/ROCm CI: `multimodal-gen-unit-test-amd` 新增 9 个用例 (`progressive 3 + sana_wm 4 + lingbot 2`), 嵌套文件启用数从 9 增至 12, 用例均为 CPU 侧 stub 级, 运行时间增量可忽略。对 CUDA: 零影响, 扁平 glob 与作业定义字节级不变。对团队: 确立了 "嵌套测试需 stub 修复时的分批启用模板", 后续 #31844、#31846 可复用同一模式; 覆盖率报告的多后端归属机制得到验证。- 风险标记: AMD 专用收集分支, 依赖上游 stub 修复, 继承 main 已有失败

关联脉络

- PR #31483 [AMD] ci: run vetted nested multimodal_gen unit tests on AMD: 直接前置 PR, 定义 `_AMD_READY_NESTED_UNIT_TESTS` 与 `_MM_GEN_FILE_BACKENDS` 机制并启用 9 个无需改动的嵌套测试文件; 本 PR 是其 body 中显式拆分的 harness stub 修复 follow-up。
- PR #31844 [AMD] ci: env-dependency skips for nested multimodal_gen unit tests: 同系列拆分 PR: 处理 LAPACK camera-pose 与 msgpack 环境依赖缺失; 本 PR body 明确将相关文件 defer 到 #31844, 且验证中确认其保持 CUDA 标签。
- PR #31846 [AMD] ci: quarantine pre-existing nested multimodal_gen unit failures: 同系列拆分 PR: 隔离 `test_realtime_webui`、`test_realtime_runtime`、`test_streaming_forward_long` 等预先存在的非 AMD 特有失败; 本 PR body 明确其文件保持 CUDA 标签。
- PR #33725 (上游) 为 GPUWorker 增加 `is_output_rank`: body 指出该 PR 在 `GPUWorker.__init__` 中新增 `is_output_rank`, 导致 `test_output_materialization.py` 的 `GPUWorker.__new__` stub 缺少该属性而抛出 `AttributeError`, 是 AMD 车道上继承自 main 的既有失败来源。