

PR #34100 完整报告

sgl-project/sglang

[Fix] Give the piecewise CUDA graph test stub an `hf\_config`

合并时间: 2026-08-08 17:04

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34100>

执行摘要

- 一句话: 修复 piecewise CUDA graph 测试 stub 缺失 hf_config 导致的 CI 失败
- 推荐动作: 值得快速浏览以理解 piecewise CUDA graph 兼容性规则如何读取 hf_config, 以及测试 stub 需要跟随生产逻辑同步更新。但整体改动直白, 无精读必要。

功能与动机

PR body 明确指出 `base-a-test-cpu` 当前在 main 上红:

`test_trtllm_mla_stays_on_breakable_and_is_disabled_by_compatibility` 抛出 `AttributeError: 'types.SimpleNamespace' object has no attribute 'hf_config'`。原因是 #32785 在 `_disable_breakable_cudagraph_if_incompatible` 的 MLA 规则中新增了 `is_deepseek_dsa(self.get_model_config().hf_config)` 项, 但该测试的 stub 只设置了 `is_multimodal_pieewise_cuda_graph_supported`。由于该文件内只有这个用例 patch 了 `use_mla_backend` 为 True, 因此只有它穿过短路逻辑触及新条款。

实现拆解

本次变更仅涉及 1 个测试文件, 实施步骤如下:

1. 定位失败根因: `test/registered/unit/configs/test_multimodal_pieewise_cuda_graph.py` 中 `test_trtllm_mla_stays_on_breakable_and_is_disabled_by_compatibility` 构造的 `SimpleNamespace` stub 只包含 `is_multimodal_pieewise_cuda_graph_supported`, 缺少 `hf_config` 属性; 而 #32785 引入的 MLA 兼容性规则会读取 `hf_config` 做 DSA 豁免判断, 导致 `AttributeError`。
2. 补齐 stub 属性: 在 `args.model_config` 的 `SimpleNamespace` 中追加 `hf_config=SimpleNamespace(architectures=["DeepseekV2ForCausalLM"])`。该架构满足“MLA 但非 DSA”的条件, 与断言期望的 `prefill.backend == Backend.DISABLED` 一致。PR body 还说明这与同文件 `test_embedding_gemma_forces_breakable_prefill` 的 stub 构建方式一致。
3. 无源码主路径改动: 本 PR 仅修复测试配置, 不涉及 `ServerArgs`、兼容性规则或任何生产代码, 因此不影响运行时行为。

关键文件:

- `test/registered/unit/configs/test_multimodal_pieewise_cuda_graph.py` (模块 配置测试; 类别 test; 类型 test-coverage; 符号 `test_trtllm_mla_stays_on_breakable_and_is_dis`

abled_by_compatibility)：唯一变更文件，修复 test_trtllm_mla_stays_on_breakable_and_is_disabled_by_compatibility 中 stub 缺失 hf_config 导致的 AttributeError，使 CI 恢复绿色。

关键符号：test_trtllm_mla_stays_on_breakable_and_is_disabled_by_compatibility

关键源码片段

test/registered/unit/configs/test_multimodal_pieewise_cuda_graph.py

唯一变更文件，修复 test_trtllm_mla_stays_on_breakable_and_is_disabled_by_compatibility 中 stub 缺失 hf_config 导致的 AttributeError，使 CI 恢复绿色。

```
# test/registered/unit/configs/test_multimodal_pieewise_cuda_graph.py

def test_trtllm_mla_stays_on_breakable_and_is_disabled_by_compatibility(self):
    args = ServerArgs(model_path="dummy")
    # The MLA rule reads hf_config to exempt DSA models, so the stub needs
    # an architecture that is MLA but not DSA.
    args.model_config = SimpleNamespace(
        is_multimodal_pieewise_cuda_graph_supported=True,
        hf_config=SimpleNamespace(architectures=["DeepseekV2ForCausalLM"]),
    )
    args.cuda_graph_config = CudaGraphConfig(
        prefill=PhaseConfig(backend=Backend.BREAKABLE)
    )
    args._cuda_graph_config_locked = set()

    with (
        patch.object(
            args,
            "_resolved_attention_backends",
            return_value=("trtllm_mla", "trtllm_mla"),
        ),
        patch.object(args, "use_mla_backend", return_value=True),
    ):
        args._apply_cuda_graph_compatibility()

    self.assertEqual(args.cuda_graph_config.prefill.backend, Backend.DISABLED)
```

评论区精华

没有 review 评论线程。PR body 中作者说明了两个关键决策：一是只有 test_trtllm_mla_stays_on_breakable_and_is_disabled_by_compatibility 会穿过短路达到新条款，因此只需修这一个 stub；二是选择 DeepseekV2ForCausalLM 作为架构，因为它代表 MLA 而非 DSA，正好对应测试断言期望的 DISABLED 后端。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低，因为变更仅限测试 stub，不触碰生产逻辑。但有两个潜在注意点：
 - 测试 stub 与真实模型配置的耦合：未来若 hf_config 结构变化（如 architectures 字段改名），该测试会再次失效，且这种失败不会被生产代码触发，需要靠 CI 捕获。
 - 该测试依赖 trtllm_mla backend 在测试环境中的可用性，若 backend 名称变更或移除，同样可能造成测试失败。
 - 影响：影响范围很小：修复了 CI 基线（base-a-test-cpu）的红灯，恢复开发者对分支状态的正确感知，提升 CI 可信度。对用户无运行时影响，对团队是低成本的维护性改进。
 - 风险标记：测试 stub 与生产配置耦合，依赖 trtllm_mla backend 可用性

关联脉络

- PR #32785 fix: avoid piecewise prefill graph for trtllm_mla: 该 PR 在 MLA 兼容性规则中引入了 is_deepseek_dsa(self.get_model_config().hf_config) 判断，直接导致本 PR 修复的测试 stub 缺失 hf_config 属性而失败，两者是因果关联。