PR #6728 完整报告

verl-project/verl

[rollout] fix: Adjust cuda graph capture sizes config logic for vLLM >= 0.11.1 compatibility

合并时间：2026-06-15 11:01

原文链接：http://prhub.com.cn/verl-project/verl/pull/6728

执行摘要

- 一句话：修复 vLLM 0.11.1+ CUDA graph 配置兼容性
- 推荐动作：值得合并，修复及时且无风险。建议后续在版本升级时增加针对 vLLM 配置兼容性的测试。设计思路（将版本兼容逻辑集中在一处条件判断）值得借鉴。

功能与动机

vLLM 0.11.1 移除了 CLI 参数 cuda_graph_sizes，改为在 compilation_config 中配置 cudagraph_capture_sizes。直接传递旧参数会引发运行时错误。PR body 引用了 vLLM 文档作为依据，并说明需要保持向后兼容。

实现拆解

1. 版本判断前置：引入 _VLLM_VERSION（此前已有定义）与 version.parse 比较。
2. 旧版本路径：当 _VLLM_VERSION <= 0.11.0 时，将 self.config.cudagraph_capture_sizes 赋值给 engine_kwargs["cuda_graph_sizes"]（原逻辑）。
3. 新版本路径：在 compilation_config 组装之后、json.dumps 之前，若 _VLLM_VERSION > 0.11.0 且存在配置，则将 self.config.cudagraph_capture_sizes 写入 compilation_config["cudagraph_capture_sizes"]。
4. 变更仅涉及一个文件（verl/workers/rollout/vllm_rollout/vllm_async_server.py），3 行新增、1 行修改，无测试配套改动。

关键文件：

- verl/workers/rollout/vllm_rollout/vllm_async_server.py（模块 rollout；类别 source；类型 core-logic；符号 launch_server）：核心变更文件，修改了 CUDA graph 配置的分支逻辑以兼容 vLLM 0.11.1+

关键符号：launch_server

关键源码片段

verl/workers/rollout/vllm_rollout/vllm_async_server.py

核心变更文件，修改了 CUDA graph 配置的分支逻辑以兼容 vLLM 0.11.1+

```
async def launch_server(self, master_address: str = None, master_port: int = None, dp_rpc_
```

```
port: int = None):
    # ... ( 省略前置逻辑 )
    engine_kwargs = {key: val for key, val in engine_kwargs.items() if val is not None}
    # ...
    # vLLM <= 0.11.0 使用顶层参数传递
    if self.config.cudagraph_capture_sizes and _VLLM_VERSION <= version.parse("0.11.0"):
        engine_kwargs["cuda_graph_sizes"] = self.config.cudagraph_capture_sizes

    self._preprocess_engine_kwargs(engine_kwargs)
    # ...
    compilation_config = engine_kwargs.pop("compilation_config", None) or {}
    # ... 省略 compilation_config 初始化与 DCP 降级逻辑
    # vLLM > 0.11.0 配置迁移到 compilation_config 内部
    if self.config.cudagraph_capture_sizes and _VLLM_VERSION > version.parse("0.11.0"):
        compilation_config["cudagraph_capture_sizes"] = self.config.cudagraph_capture_sizes

    compilation_config = json.dumps(compilation_config)
    # ... 后续组装 args 并传给 vLLM 服务器
```

评论区精华

Reviewer Luosuu 快速批准（LGTM）。ConanZH429 指出 6 个 CI 失败均与本次改动无关（Ray 连接失败、OOM 等 flaky 问题），要求重跑或合并。无其他设计争论。

- 暂无高价值评论线程

风险与影响

- 风险：风险低。改动范围极小，版本判断基于已有 _VLLM_VERSION 变量，逻辑清晰。但未增加单元测试，未来 vLLM 进一步变更配置结构时可能再次不兼容。依赖 _VLLM_VERSION 定义的准确性。
- 影响：直接影响所有使用 vLLM rollout 的 verl 用户。对 vLLM <= 0.11.0 用户无行为变化；对 vLLM >= 0.11.1 用户修复了 CUDA graph 配置失败导致的崩溃。影响范围局限在服务器启动阶段的配置逻辑。
- 风险标记：缺少测试覆盖，依赖 vLLM 版本号

关联脉络

- PR #6661 [rollout, vllm] fix: preserve MTP drafter weights during hybrid sleep: 同为 rollout/vllm 模块的兼容性修复，改动同一文件 vllm_async_server.py
- PR #6630 [rollout] fix: propagate ignore_eos to vLLM sampling params: 同为 rollout/vllm 模块的配置传递修复，同样改动 vllm_async_server.py