

PR #29591 完整报告

sgl-project/sglang

[misc] Use `--cuda-graph-max-bs-decode` in tests, examples, and docs

合并时间: 2026-06-29 09:38

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29591>

执行摘要

- 一句话: 统一替换 `--cuda-graph-max-bs` 为 `--cuda-graph-max-bs-decode`
- 推荐动作: 建议立即合入, 属于有益的技术债务清理。作者提供的验证脚本可作为类似重命名操作的参考。

功能与动机

PR body 指出 `--cuda-graph-max-bs` 已被标记弃用, 但内部代码仍广泛使用, 导致弃用警告频繁输出。为规范参数使用并减少噪音, 需统一替换为标准参数名。

实现拆解

步骤

1. 识别替换范围: 涵盖所有 CI 测试文件、示例脚本和文档配置片段, 这些地方硬编码了 `--cuda-graph-max-bs` 或 `cuda_graph_max_bs`。
2. 自动执行替换: 通过 `transform.sh` 脚本 (参见 [gist](#)) 执行确定性正则替换, 将 `cuda_graph_max_bs` (无短横) 和 `--cuda-graph-max-bs` 一并更新为带 `-decode` 后缀的形式。
3. 等效性验证: `verify.sh` 脚本证明 136 个文件在忽略 `-decode` 后缀后与原始 diff 字节一致; 另 4 个文件因 black 格式化调整了换行但 AST 等价。
4. 提交合并: 单 commit 打包变更, 确保原子性。

关键文件:

- `docs_new/src/snippets/autoregressive/deepseek-r1-advanced-deployment.jsx` (模块 文档配置; 类别 `source`; 类型 `core-logic`): DeepSeek R1 高级部署配置片段, 用户常参考, 修改覆盖所有参数实例, 是本次替换的代表性文档文件。
- `examples/runtime/engine/offline_batch_inference_eagle.py` (模块 示例代码; 类别 `source`; 类型 `core-logic`): EAGLE 示例推理脚本, 展示了 Python API 中参数名的正确用法, 是用户直接复制的模板。
- `test/registered/moe/test_hybrid_dp_ep_tp_mtp.py` (模块 测试用例; 类别 `test`; 类型 `test-coverage`): MoE 混合并行测试, 大量重复使用该参数, 是测试文件中最典型的代表。

关键符号: 未识别

关键源码片段

[docs_new/src/snippets/autoregressive/deepseek-r1-advanced-deployment.js](#)
x

DeepSeek R1 高级部署配置片段，用户常参考，修改覆盖所有参数实例，是本次替换的代表性文档文件。

```
/* B200 + FP4 低延迟配置片段：参数名已更新 */
{
  "hardware": "b200",
  "quantization": "fp4",
  "gpu_count": 4,
  "scenario": "low-latency",
  "parameters": {
    "model_path": "nvidia/DeepSeek-R1-0528-FP4-v2",
    "tensor_parallel_size": 4,
    "cuda_graph_max_bs_decode": 256, // 原为 cuda_graph_max_bs
    "max_running_requests": 256,
    "mem_fraction_static": 0.85,
    "ep_size": 4,
    "scheduler_recv_interval": 10,
    "enable_symm_mem": true,
    "stream_interval": 10
  }
}
```

[test/registered/moe/test_hybrid_dp_ep_tp_mtp.py](#)

MoE 混合并行测试，大量重复使用该参数，是测试文件中最典型的代表。

```
# 部分 setUpClass 中的参数列表片段
[
  "--tp", "8",
  "--moe-a2a-backend", "deepep",
  "--cuda-graph-max-bs-decode", "128", # 原为 --cuda-graph-max-bs
  # ...
]
```

评论区精华

本 PR 无人工 review 讨论，作者在 PR body 中提供了自动脚本和验证方法以确保透明。bot 评论仅涉及配额与文档预览。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。替换经过脚本自动化和字节级 /AST 级双重验证，确保无逻辑漂移。唯一潜在风险是用户参考了旧参数名，但因旧别名保留，功能不受影响。变更不涉及运行时路径，不影响性能或正确性。

- 影响：影响范围：140 个文件，涵盖 CI 测试、示例和文档。影响程度：低，不改变行为，仅消除弃用警告。用户运行测试或参考示例时不再看到警告；文档同步更新避免误导；社区用户若使用旧别名仍正常工作。
- 风险标记：无逻辑变更，验证脚本完备，向后兼容

关联脉络

- 暂无明显关联 PR