

PR #30016 完整报告

sgl-project/sglang

[diffusion] feat: performance_mode=speed enables torch.compile by default

合并时间: 2026-07-03 19:12

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30016>

执行摘要

- 一句话: diffusion speed 模式默认开启 torch.compile
- 推荐动作: 值得合并, 逻辑简洁且测试覆盖充分。CI 状态显示 Extra 测试失败, 需确认是否与此变更相关。

功能与动机

PR 指出 `--performance-mode speed` 虽配置了 GPU 常驻执行 (关闭 offload、启用 FSDP/CFG-parallel), 但除非用户额外传递 `--enable-torch-compile`, 否则仍运行在 eager 模式。跨框架基准审计发现多个速度意图部署静默地以 eager 模式运行 DiT。因此“speed”应默认最快。

实现拆解

1. 核心逻辑修改: 在 `server_args_auto_tune.py` 的 `adjust_based_on_performance_mode` 方法中, 当 `performance_mode == "speed"` 时, 检查 `enable_torch_compile` 是否为 `False` 且未被用户显式设置 (通过 `is_arg_explicitly_set("enable_torch_compile")`)。若满足条件, 则将其设为 `True` 并记录日志。
2. 显式覆盖优先: 若用户已显式传递 `--enable-torch-compile false`, 则跳过默认开启, 保留其选择。
3. 测试配套: 在 `test_server_args.py` 中添加三个单元测试: 验证 speed 模式默认开启 compile、验证显式关闭有效、验证 auto 模式不开启 compile。

关键文件:

- `python/sglang/multimodal_gen/runtime/server_args_auto_tune.py` (模块 自动调优; 类别 source; 类型 core-logic; 符号 `adjust_based_on_performance_mode`): 核心逻辑变更: 在 speed 模式分支中自动启用 torch.compile
- `python/sglang/multimodal_gen/test/unit/test_server_args.py` (模块 测试; 类别 test; 类型 test-coverage; 符号 `test_speed_mode_enables_torch_compile_by_default`, `test_speed_mode_preserves_explicit_torch_compile_off`, `test_auto_mode_leaves_torch_compile_off`): 添加三个单元测试覆盖新行为

关键符号: `adjust_based_on_performance_mode`

关键源码片段

python/sglang/multimodal_gen/runtime/server_args_auto_tune.py

核心逻辑变更：在 speed 模式分支中自动启用 torch.compile

```
def adjust_based_on_performance_mode(self) -> None:
    """Adjust the server args based on the performance mode"""
    args = self.server_args
    args.performance_mode = self._normalize_performance_mode()

    if current_platform.is_cpu():
        return

    if args.performance_mode == "speed":
        logger.info("Applying performance_mode=speed")
        # speed 模式默认开启 torch.compile, 除非用户显式关闭
        if not args.enable_torch_compile and not args.is_arg_explicitly_set(
            "enable_torch_compile"
        ):
            args.enable_torch_compile = True
            logger.info(
                "performance_mode=speed enables torch.compile "
                "(pass --enable-torch-compile false to opt out)"
            )
        if args.num_gpus >= 2 and self._can_apply_fsdp_policy(
            require_memory_headroom=False
        ):
            self._set_gpu_resident_defaults(use_fsdp=True)
            self._enable_cfg_parallel_if_supported()
        else:
            self._set_gpu_resident_defaults(use_fsdp=False)
        return
    # ... memory 分支保持不变
```

python/sglang/multimodal_gen/test/unit/test_server_args.py

添加三个单元测试覆盖新行为

```
def test_speed_mode_enables_torch_compile_by_default(self):
    # 仅设置 performance_mode="speed", 期望 enable_torch_compile 为 True
    args = self._from_dict_with_pipeline_config(
        QwenImagePipelineConfig(),
        kwargs={
            "model_path": "Qwen/Qwen-Image",
            "performance_mode": "speed",
        },
    )
    self.assertTrue(args.enable_torch_compile)

def test_speed_mode_preserves_explicit_torch_compile_off(self):
    # 显式关闭 compile, 期望保持 False
    args = self._from_dict_with_pipeline_config(
```

```

        QwenImagePipelineConfig(),
        kwargs={
            "model_path": "Qwen/Qwen-Image",
            "performance_mode": "speed",
            "enable_torch_compile": False,
        },
    )
    self.assertFalse(args.enable_torch_compile)

def test_auto_mode_leaves_torch_compile_off(self):
    # auto 模式不应开启 compile
    args = self._from_dict_with_pipeline_config(
        QwenImagePipelineConfig(),
        kwargs={
            "model_path": "Qwen/Qwen-Image",
            "performance_mode": "auto",
        },
    )
    self.assertFalse(args.enable_torch_compile)

```

评论区精华

仅有 [gemini-code-assist\[bot\]](#) 的自动回复，无实质讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。若用户期望 speed 模式但不需要 compile（如短步数模型），可通过显式 `--enable-torch-compile false` 关闭，且日志已注明。但新增默认行为可能导致某些用户意外遇到 compile 开销，需关注相关 issue。
- 影响：影响扩散模型用户，特别是使用 `--performance-mode speed` 的用户。默认行为变化提升了性能，但可能对短步数模型有负作用，用户需了解 opt-out 机制。
- 风险标记：额外 CI 失败需确认

关联脉络

- 暂无明显关联 PR