

PR #34174 完整报告

sgl-project/sglang

[diffusion] BCG: auto-capture the default warmup resolution instead of hard-requiring --warmup-resolutions (H200 SANA denoise 0.73->0.457 s with a single flag)

合并时间: 2026-08-10 08:37

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34174>

执行摘要

- 一句话: BCG 自动捕获默认 warmup 分辨率, 单 flag 即可启用
- 推荐动作: 值得精读: 改动虽小, 但展示了“复用既有默认值解析 + 失败降级 + 日志提示”的组合设计。建议关注两点: 一是自动种入的默认分辨率是否总符合部署预期; 二是补充覆盖 _default_bcg_warmup_resolution 成功与失败分支的集成测试, 避免后续回归。

功能与动机

PR body 指出: BCG 图按分辨率在 warmup 时捕获, 因此分辨率必须预先声明, 但要求用户总是写全 --warmup-resolutions 让该 flag 很难开启, 也导致“只翻转这个开关的 eager 基准 / 部署”从未真正走到 BCG 路径。核心诉求是让 --enable-breakable-cuda-graph 可独立使用, 并复用 warmup builder 已有的默认分辨率选择能力。

实现拆解

1. 放宽校验: _validate_breakable_cuda_graph 将 warmup_resolutions 为空时的 ValueError 改为 logger.info 提示, 保留 bcg_text_buckets 的正整数校验, 避免其它错误被吞掉。
2. 自动种入默认值: _adjust_breakable_cuda_graph_support 在模型受 BCG 支持且 warmup_resolutions 未设置时, 调用新方法 _default_bcg_warmup_resolution 生成默认分辨率并写入 self.warmup_resolutions。
3. 默认分辨率推导: 新增 _default_bcg_warmup_resolution, 延迟导入 warmup_request_builder 的 get_model_sampling_defaults 与 _resolve_default_warmup_resolution, 以 server_based_warmup=True 解析模型默认分辨率; 解析失败时记录 warning 并保持 warmup_resolutions 为空, 即不捕获任何图、全程 eager。
4. 测试配套: python/sglang/multimodal_gen/test/unit/test_server_args.py 新增 test_breakable_cuda_graph_allows_unset_resolutions, 用裸 ServerArgs 实例验证 _validate_breakable_cuda_graph 不再抛异常; 该测试只覆盖校验路径, 未覆盖自动种入的集成行为。

关键文件:

- python/sglang/multimodal_gen/runtime/server_args/server_args.py (模块 参数配置; 类别 source; 类型 core-logic; 符号 _default_bcg_warmup_resolution, _validate_breakable_cuda_graph, _adjust_breakable_cuda_graph_support) : 核心变更所在: BCG 分辨率校验从硬错误改为软提示, 并在启动调整阶段自动种入模型默认 warmup 分辨率。
- python/sglang/multimodal_gen/test/unit/test_server_args.py (模块 单元测试; 类别 test; 类型 test-coverage; 符号 test_breakable_cuda_graph_allows_unset_resolutions) : 新增回归测试, 验证未声明 --warmup-resolutions 时 _validate_breakable_cuda_graph 不再抛异常。

关键符号: _default_bcg_warmup_resolution, _validate_breakable_cuda_graph, _adjust_breakable_cuda_graph_support

关键源码片段

python/sglang/multimodal_gen/runtime/server_args/server_args.py

核心变更所在: BCG 分辨率校验从硬错误改为软提示, 并在启动调整阶段自动种入模型默认 warmup 分辨率。

```
# python/sglang/multimodal_gen/runtime/server_args/server_args.py
```

```
def _validate_breakable_cuda_graph(self):
    # BCG 启动校验: 分辨率未声明时不再硬报错, 改为提示日志。
    if not self.enable_breakable_cuda_graph:
        return
    if not self.warmup_resolutions:
        # BCG 图按分辨率捕获, 仅在 warmup 阶段生成; 运行期不重新捕获。
        # 未显式声明时, 由 _adjust_breakable_cuda_graph_support 在启动期
        # 种入模型默认分辨率, 其它分辨率请求回退 eager。
        logger.info(
            '[Diffusion BCG] --warmup-resolutions unset; capturing the '
            'model default warmup resolution. Requests at other '
            'resolutions run eager.'
        )
    if self.bcg_text_buckets is not None and not any(
        int(b) > 0 for b in self.bcg_text_buckets
    ):
        raise ValueError(
            '--bcg-text-buckets must contain at least one positive integer.'
        )
```

```
def _default_bcg_warmup_resolution(self):
    # 为未声明 --warmup-resolutions 的 BCG 场景种入模型默认分辨率。
    # 延迟导入避免 server_args 与 warmup 构建器产生循环依赖。
    from sglang.multimodal_gen.runtime.warmup_request_builder import (
        _resolve_default_warmup_resolution,
        get_model_sampling_defaults,
    )
```

```

try:
    sampling_defaults = get_model_sampling_defaults(self)
    width, height = _resolve_default_warmup_resolution(
        self, sampling_defaults, server_based_warmup=True
    )
except Exception as exc: # pragma: no cover - defensive
    # 推导失败时不捕获任何图，服务全程 eager，并记录告警避免静默。
    logger.warning(
        '[Diffusion BCG] could not derive a default warmup resolution '
        '(%s); no graph will be captured and serving runs eager.',
        exc,
    )
    return
self.warmup_resolutions = [f'{width}x{height}']
logger.info(
    '[Diffusion BCG] --warmup-resolutions unset; capturing the model '
    'default %dx%d. Requests at other resolutions run eager.',
    width,
    height,
)

```

评论区精华

本 PR 全程无 review 评论。作者 BBuf 在 PR 内留言确认 CI 全绿：required 检查 14/14 [call-multimodal-gen-tests](#) SUCCESS, [finish/lint/gate](#) SUCCESS；剩余红灯为已知非必需 AMD/NPU/finish-aggregator 车道，合并的 #34008/#34085 同样如此。

- CI 状态确认 (other): CI 满足合并条件，PR 已由作者合并。

风险与影响

- 风险：

1. 行为变更：未声明 `--warmup-resolutions` 时启动不再报错，可能掩盖用户配置遗漏；非默认分辨率请求静默回退 `eager`，性能差异可能未被察觉。
2. 默认分辨率推导依赖 `get_model_sampling_defaults` 与 `_resolve_default_warmup_resolution`，异常时被 `except` 捕获并降级为不捕获任何图，服务仍可运行但失去 BCG 加速，用户可能只看到 `warning`。
3. 测试仅覆盖 `_validate_breakable_cuda_graph` 路径，未覆盖 `_adjust_breakable_cuda_graph_support` 自动种入与推导失败的降级路径。- 影响：影响范围限于 `multimodal_gen` 的 `diffusion` 服务启动配置阶段，运行时逻辑不变。用户侧：diffusion BCG 开箱可用，H200 + SANA 场景 `denoise` 0.73s→0.457s（约 37% 提升），`e2e` 0.90s→0.636s；benchmark 侧：后续 `eager` 与 BCG 对比可直接单 `flag` 开启。团队侧：为 BCG 默认化与统一基准铺路。- 风险标记：未声明分辨率不再报错（行为变更），非默认分辨率静默回退 `eager`，推导失败时无图捕获仅告警，测试未覆盖自动种入集成路径

关联脉络

- PR #33702 [diffusion] Add Sol-Attn sparse attention backend for diffusion: 同属 diffusion 性能优化功能线（引入新 attention 后端提升 H3 去噪速度），本 PR 降低 BCG 使用门槛，两者无文件交集但方向一致。