PR #35724 完整报告

sgl-project/sglang

[diffusion] Enable LongCat breakable CUDA graphs

合并时间：2026-08-21 17:59

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/35724

执行摘要

- 一句话：为 LongCat-Image 启用可中断 CUDA Graph，端到端提速约 3.4%
- 推荐动作：值得精读。核心价值不在改动量，而在于 " 模型固定 prompt 形状时用 pass-through padder 绕过通用 bucket padding" 的设计决策，以及 BCG 按分辨率捕获、未命中回退 *eager* 的运行时契约。对有意给 diffusion 模型接入 BCG 的工程师，longcat_image.py 与 server_args.py 的白名单机制是最小可复用范例。

功能与动机

LongCat 的原生 *eager* DiT *trace* 在 *attention* 调用之间 launch 开销密集，已有 BCG runner 可将这些区域捕获成 31 个图分段，但 LongCat 此前未进入模型与 pipeline 白名单。PR body 明确："LongCat always sends the full 512-token prompt body to the DiT. A small model-specific pass-through padder therefore preserves one reusable signature for short and long raw prompts instead of applying generic prompt padding." 即通用 bucket padding 只会为固定 512-token 输入生成更大且未使用的 *graph* 签名，因此需要直通 padder 保留单一可复用签名。

实现拆解

1. 新增 LongCat 专用 padder 模块：新建 python/sglang/multimodal_gen/runtime/breakable_cuda_graph/model_padders/longcat_image.py（+33 行），定义 is_longcat_image_transformer 谓词（类名模糊匹配 longcatimage 且 call_kwargs 含 encoder_hidden_states 与 txt_ids 键）与 keep_longcat_prompt_shape 直通函数（原样返回 call_kwargs，不做任何 bucket padding），并在模块加载时通过 register_prompt_padder 注册。短长 raw prompt 因此复用同一个 512-token 捕获签名。
2. 注册入口打通：breakable_cuda_graph/prompt_padding.py 的 _ensure_model_padders_registered 补入 longcat_image 导入（+1 行），使 padder 在 BCG 首次使用时自动注册，无需改动选择逻辑。
3. 服务参数白名单扩展：server_args.py 在 BREAKABLE_CUDA_GRAPH_SUPPORTED_MODEL_IDS 中加入 meituan-longcat/longcat-image，在 BREAKABLE_CUDA_GRAPH_SUPPORTED_PIPELINE_CONFIGS 中加入 LongCatImagePipelineConfig，并同步 _adjust_breakable_cuda_graph_support 的禁用警告文案，使 --enable-breakable-cuda-graph 对 LongCat 部署直接生效。

4. 测试配套：test_diffusion_bcg_padding.py（+27 行）新增 LongCatImageTransformer2DModel 桩模型、setUp 实例与 test_longcat_keeps_its_fixed_512_token_prompt_shape，断言 _bcg_pad_prompt_kwargs 返回原对象（assertIs）、encoder_hidden_states 保持 (1, 512, 3584)、txt_ids 保持 (512, 3)，并把 LongCat 模型 ID 与 pipeline config 纳入 " 已注册为 BCG 支持 " 的参数化测试。
5. 文档与技能库更新：docs/docs/sglang-diffusion/api/cli.mdx 给出 1024x1024 部署命令（含 --warmup-resolutions 与 --enable-torch-compile false）；.claude/skills/ 下两个 skill 文档记录 LongCat 固定 512-token 签名机制与 H200/B300 实测数据，避免性能工程师重复提案。

关键文件：

- python/sglang/multimodal_gen/runtime/breakable_cuda_graph/model_padders/longcat_image.py（模块 BCG 填充；类别 source；类型 data-contract；符号 is_longcat_image_transformer, keep_longcat_prompt_shape）：核心新增文件：定义 LongCat 专用直通 padder 并注册，是保持固定 512-token 签名、避免通用 bucket 膨胀的关键逻辑。
- python/sglang/multimodal_gen/test/unit/test_diffusion_bcg_padding.py（模块 单元测试；类别 test；类型 test-coverage；符号 LongCatImageTransformer2DModel, test_longcat_keeps_its_fixed_512_token_prompt_shape）：测试配套：新增 LongCat 桩模型与固定 512-token prompt shape 的直通断言测试，并把 LongCat 纳入 BCG 注册白名单参数化测试。
- python/sglang/multimodal_gen/runtime/server_args/server_args.py（模块 服务参数；类别 source；类型 core-logic）：BCG 白名单与启停逻辑所在：加入 LongCat 模型 ID 与 pipeline config，是启用能力生效的入口。
- python/sglang/multimodal_gen/runtime/breakable_cuda_graph/prompt_padding.py（模块 BCG 注册；类别 source；类型 core-logic）：padder 注册入口：补入 longcat_image 导入，使新增 padder 在 BCG 首次使用时自动注册。
- docs/docs/sglang-diffusion/api/cli.mdx（模块 CLI 文档；类别 docs；类型 documentation）：用户文档：给出 LongCat-Image 启用 BCG 的部署命令示例与固定 512-token 说明。
- python/sglang/multimodal_gen/.claude/skills/sglang-diffusion-benchmark-profile/existing-fast-paths.md（模块 基准文档；类别 docs；类型 documentation）：基准 skill 文档：记录 LongCat 固定 512-token 签名机制与 H200/B300 实测数据，避免重复提案。
- python/sglang/multimodal_gen/.claude/skills/sglang-diffusion-performance/SKILL.md（模块 技能文档；类别 docs；类型 documentation）：性能 skill 文档：同步 LongCat BCG 支持状态，保持技能库与实现一致。

关键符号：is_longcat_image_transformer, keep_longcat_prompt_shape

关键源码片段

python/sglang/multimodal_gen/runtime/breakable_cuda_graph/model_padders/longcat_image.py

核心新增文件：定义 LongCat 专用直通 padder 并注册，是保持固定 512-token 签名、避免通用 bucket 膨胀的关键逻辑。

```
# Copyright 2023-2026 SGLang Team
# Licensed under the Apache License, Version 2.0
# ==============================================================================
"""LongCat-Image breakable CUDA graph (BCG) prompt handling."""

from __future__ import annotations

from typing import Any

from sglang.multimodal_gen.runtime.breakable_cuda_graph import (
    prompt_padding as bcg_utils,
)


def is_longcat_image_transformer(current_model: Any, call_kwargs: dict) -> bool:
    # 仅当当前 DiT 是 LongCat-Image 且调用参数同时携带
    # encoder_hidden_states 与 txt_ids 时，才走 LongCat 专用直通 padder。
    return (
        bcg_utils.transformer_class_name_matches(current_model, "longcatimage")
        and "encoder_hidden_states" in call_kwargs
        and "txt_ids" in call_kwargs
    )


def keep_longcat_prompt_shape(
    call_kwargs: dict, _current_model: Any, _buckets: tuple[int, ...]
) -> dict:
    # LongCat 始终向 DiT 提供完整的 512-token prompt body，
    # 因此通用 bucket padding 只会生成更大且用不到的 graph 签名。
    # 直接原样返回 call_kwargs，让短长 prompt 复用同一个捕获签名。
    return call_kwargs


bcg_utils.register_prompt_padder(
    is_longcat_image_transformer, keep_longcat_prompt_shape
)
```

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

BCG 白名单与启停逻辑所在：加入 LongCat 模型 ID 与 pipeline config，是启用能力生效的入口。

```
# Prompt 序列长度桶：提示条件会被 padding 到能容纳它的最小桶，
```

```
# 使不同长度 prompt 共享同一捕获 graph。
DEFAULT_BCG_TEXT_BUCKETS = (64, 128, 256, 512, 1024)

# BCG 可用模型 ID 白名单（本 PR 新增 meituan-longcat/longcat-image）
BREAKABLE_CUDA_GRAPH_SUPPORTED_MODEL_IDS = frozenset({
    "comfy-org/ideogram-4",
    "efficient-large-model/sana1.5_1.6b_1024px_diffusers",
    "fal/ideogram-v4-fast",
    "fal/ideogram-v4-instant",
    "glm-image",
    "ideogram-4",
    "ideogram-4-fp8",
    "ideogram-4-nf4",
    "lightricks/ltx-2",
    "lightricks/ltx-2.3",
    "meituan-longcat/longcat-image", # 新增：LongCat-Image 支持 BCG
    "minimax-h3",
    "minimaxai/minimax-h3",
    "qwen/qwen-image",
    "qwen/qwen-image-2512",
    "tongyi-mai/z-image",
    "tongyi-mai/z-image-turbo",
    "zai-org/glm-image",
    "z-image",
    "z-image-turbo",
})

# 对应 pipeline config 白名单（新增 LongCatImagePipelineConfig）
BREAKABLE_CUDA_GRAPH_SUPPORTED_PIPELINE_CONFIGS = frozenset({
    "GlmImagePipelineConfig",
    "Ideogram4PipelineConfig",
    "LTX2PipelineConfig",
    "LTX23PipelineConfig",
    "LongCatImagePipelineConfig",
    "MiniMaxH3PipelineConfig",
    "QwenImagePipelineConfig",
    "SanaPipelineConfig",
    "ZImagePipelineConfig",
})



def _adjust_breakable_cuda_graph_support(self):
    if not self.enable_breakable_cuda_graph:
        return

    # 只有 pipeline config 与模型 ID 同时命中白名单时才保留 BCG。
    pipeline_config = getattr(self, "pipeline_config", None)
    pipeline_config_name = type(pipeline_config).__name__
    if (
```

```
        pipeline_config_name in BREAKABLE_CUDA_GRAPH_SUPPORTED_PIPELINE_CONFIGS
        and self._is_breakable_cuda_graph_supported_model()
    ):
        if not self.warmup_resolutions:
            # BCG 图按分辨率捕获；未显式声明时用模型默认 warmup 分辨率。
            self._default_bcg_warmup_resolution()
        return

    logger.warning(
        "[Diffusion BCG] disabled for %s: only Ideogram-4, "
        "Lightricks/LTX-2, LongCat-Image, MiniMax-H3, "
        "Qwen/Qwen-Image, Qwen/Qwen-Image-2512, SANA1.5, "
        "Tongyi-MAI/Z-Image/Z-Image-Turbo, and zai-org/GLM-Image are "
        "currently supported.",
        pipeline_config_name,
    )
    self.enable_breakable_cuda_graph = False
```

评论区精华

本 PR 没有任何 review 评论（review_comments_count=0），公开讨论为空。关键设计权衡集中在 PR body 与代码注释中：其一，LongCat 固定 512-token prompt body，通用 bucket padding 只会产生更大且未使用的 graph 签名，因此选择模型级直通 padder；其二，BCG 仍为 opt-in 且按分辨率生效，未声明 warmup 分辨率时回退 eager 且不重新捕获；其三，显存取舍明确——H200 上 BCG 峰值 33.4 GB 对 eager 31.7 GB，换取约 2.3% denoise step 提速，三个不同 prompt 长度全部复用同一个 31 段 graph 且输出像素级一致。

- 暂无高价值评论线程

风险与影响

- 风险：显存峰值上升：H200 实测约 +1.7 GB（33.4 vs 31.7 GB），另有一次性图捕获开销（B300 约 16.43 s），低显存部署需评估。依赖固定 512-token 假设：keep_longcat_prompt_shape 直接返回 call_kwargs，若未来 LongCat 权重或 tokenizer 使 prompt 长度可变，直通 padder 将失去签名复用价值（当前已有测试固化该假设）。分辨率绑定：BCG 图按分辨率捕获，未声明 --warmup-resolutions 的分辨率请求会静默回退 eager（见 server_args.py _adjust_breakable_cuda_graph_support），用户可能误判加速生效。验证范围有限：仅覆盖 1024x1024、BF16、50 steps 单卡（H200/B300），多分辨率、多卡与其它 dtype 行为未验证。回归面较小：默认路径完全不受影响，风险集中在启用 BCG 的 LongCat 部署。
- 影响：用户侧：LongCat-Image 部署者可通过 --enable-breakable-cuda-graph 获得约 3.4% 端到端（B300）、约 11.9% denoise stage 提速（H200 约 2.3% per step），输出与 eager 像素级一致，BCG 为 opt-in 不影响默认行为。系统侧：BCG padder 注册体系新增 " 直通型 " 成员，展示了一种避免通用 bucket 膨胀签名的模型适配模式，为后续模型接入提供模板。团队侧：文档与 skill 更新降低性能优化重复提案概率；测试扩展保证白名单与注册机制的回归安全。

- 风险标记：显存峰值上升约 1.7 GB，依赖固定的 512-token 形状假设，未预热分辨率静默回退 *eager*，缺少多分辨率 / 多卡验证

关联脉络

- PR #35728 [diffusion] Accelerate SANA-Video linear attention in quality=high: 同属 diffusion 性能优化线，都在 multimodal_gen 运行时内添加 fast path，与本 PR 构成 diffusion 加速家族的一部分。
- PR #35701 [diffusion] feat: let offloaded weights stay on the checkpoint mapping: 同为 diffusion 内存 / 性能优化（offload 与 BCG 都服务低显存场景），offload 机制可用于对冲 BCG 的峰值显存上升。
- PR #34247 [Docs] Standardize diffusion cookbook model pages: 同属 diffusion 文档标准化脉络，本 PR 在 cli.mdx 增加 LongCat 部署命令，延续该方向。