

PR #29688 完整报告

sgl-project/sglang

[diffusion][cache-dit] support Krea-2 + run-driven `has\_separate\_cfg`

合并时间: 2026-06-30 23:22

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29688>

执行摘要

- 一句话: 支持 Krea-2 Cache-DiT 加速并动态检测 CFG 模式
- 推荐动作: 推荐阅读此 PR, 它展示了如何将一个新模型家族 (Krea-2) 优雅地接入缓存系统, 并通过运行时参数解决模型变体间的 CFG 差异问题。设计决策 (动态探测 vs 静态注册) 值得借鉴, 尤其是解决蒸馏模型缓存无效的通用思路。

功能与动机

PR 要解决两个问题: 1) 为 Krea-2 (Turbo 和 Raw) 提供 Cache-DiT 支持; 2) 修正之前将所有模型变体强制赋予 `has_separate_cfg=True` 的硬编码设计。例如 `baidu/ERNIE-Image-Turbo` 和 `krea/Krea-2-Turbo` 不使用 CFG, 但硬编码为 True 会导致 `SGLANG_CACHE_DIT_ENABLED=true` 对这些蒸馏模型几乎不生效 (实测仅 1.01x 加速)。

实现拆解

1. 扩展自定义 BlockAdapter 注册表: 在 `cache_dit_integration.py` 的 `_CUSTOM_BLOCK_ADAPTER_SPECS` 中添加 `Krea2Transformer2DModel`, 指定其 `block` 属性名为 `transformer_blocks`, `forward` 模式为 `Pattern_3`。
2. 移除静态 `has_separate_cfg` 字段: 将 `_CUSTOM_BLOCK_ADAPTER_SPECS` 的值从三元组变为二元组, 不再硬编码 CFG 模式; 改为在 `_build_custom_block_adapter` 和 `enable_cache_on_transformer` 函数签名中增加 `has_separate_cfg: bool` 参数, 由调用者传入。
3. 在去噪阶段注入运行时 CFG 状态: 在 `denoising.py` 的 `_maybe_enable_cache_dit` 中, 调用 `enable_cache_on_transformer` 时传入 `has_separate_cfg=batch.do_classifier_free_guidance`, 实现每批次动态适配。
4. 统一 Krea-2 SingleStreamBlock 前向参数名: 将 `krea2.py` 中 `SingleStreamBlock.forward` 的参数 `x` 重命名为 `hidden_states`, 与 `cache-dit` 的 `Pattern_3` 期望命名一致。
5. 补充文档与测试: 在 Krea-2 的 `cookbook` 中新增 Cache-DiT 使用说明及性能数据; 添加单元测试 `test_has_separate_cfg_follows_runtime` 验证 `has_separate_cfg` 正确随运行时参数变化。

关键文件:

- `python/sglang/multimodal_gen/runtime/cache/cache_dit_integration.py` (模块 缓存层; 类别 source; 类型 core-logic; 符号 `_CUSTOM_BLOCK_ADAPTER_SPECS`, `_build_custom_block_adapter`, `enable_cache_on_transformer`): 核心变更: 修改自定义 BlockAdapter 注册表格式, 移除静态 `has_separate_cfg`, 改为运行时参数; 添加 Krea-2 注册; 修改 `_build_custom_block_adapter` 和 `enable_cache_on_transformer` 签名。
- `python/sglang/multimodal_gen/runtime/models/dits/krea2.py` (模块 模型适配; 类别 source; 类型 data-contract; 符号 `SingleStreamBlock.forward`): Krea-2 模型定义: 修改 `SingleStreamBlock.forward` 的参数名 `x` 为 `hidden_states`, 与 `cache-dit` 的 `Pattern_3` 一致; 新增 `Krea2Transformer2DModel` 类 (已存在, 但相关适配已完成)。
- `python/sglang/multimodal_gen/runtime/pipelines_core/stages/denoising.py` (模块 去噪阶段; 类别 source; 类型 core-logic; 符号 `_maybe_enable_cache_dit`): 去噪阶段调用 `enable_cache_on_transformer` 时传入 `has_separate_cfg=batch.do_classifier_free_guidance`, 实现运行时 CFG 感知。
- `python/sglang/multimodal_gen/test/unit/test_cache_dit_integration.py` (模块 测试; 类别 test; 类型 test-coverage; 符号 `test_has_separate_cfg_follows_runtime`, `test_builds_adapter_for_registered_class`): 新增 `test_has_separate_cfg_follows_runtime` 测试, 验证 Krea-2 Raw 和 Turbo 使用不同 `has_separate_cfg` 值时 adapter 正确设置; 同时修改已有测试以适配新函数签名。
- `docs_new/cookbook/diffusion/Krea/Krea-2.mdx` (模块 文档; 类别 other; 类型 documentation): 新增 Cache-DiT 使用文档, 包含基本用法、性能表格和高级参数说明。

关键符号: `_build_custom_block_adapter`, `enable_cache_on_transformer`, `SingleStreamBlock.forward`, `_maybe_enable_cache_dit`

关键源码片段

`python/sglang/multimodal_gen/runtime/cache/cache_dit_integration.py`

核心变更: 修改自定义 BlockAdapter 注册表格式, 移除静态 `has_separate_cfg`, 改为运行时参数; 添加 Krea-2 注册; 修改 `_build_custom_block_adapter` 和 `enable_cache_on_transformer` 签名。

```
def _build_custom_block_adapter(
    transformer: torch.nn.Module,
    has_separate_cfg: bool = False, # 新增参数, 由调用者 (去噪阶段) 根据 batch 的 CFG
    模式传入
) -> Optional[BlockAdapter]:
    """为未注册到 cache-dit 的模型手动构建 BlockAdapter。"""
    spec = _CUSTOM_BLOCK_ADAPTER_SPECS.get(transformer.__class__.__name__)
    if spec is None:
        return None
    blocks_attr, forward_pattern = spec # 原三元组去掉 has_separate_cfg
    blocks = getattr(transformer, blocks_attr, None)
    if blocks is None:
        raise ValueError(...)
    return BlockAdapter(
        transformer=transformer,
```

```

        blocks=blocks,
        forward_pattern=forward_pattern,
        has_separate_cfg=has_separate_cfg, # 现在从参数动态传入
    )

```

python/sglang/multimodal_gen/runtime/models/dits/krea2.py

Krea-2 模型定义：修改 SingleStreamBlock.forward 的参数名 x 为 hidden_states，与 cache-dit 的 Pattern_3 一致；新增 Krea2Transformer2DModel 类（已存在，但相关适配已完成）。

```

class SingleStreamBlock(nn.Module):
    def forward(
        self,
        hidden_states: Tensor, # 原参数名为 x, 现改为 hidden_states, 与 cache-dit 的 Pattern_3
        预期匹配
        vec: Tensor,
        freqs: Tensor,
        key_mask: Tensor | None = None,
        mask_meta: dict | None = None,
    ) -> Tensor:
        # ... 内部所有 x 引用改为 hidden_states
        hidden_states = hidden_states + pregate * self.attn(...)
        hidden_states = hidden_states + postgate * self.ff(...)
        return hidden_states

```

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

新增 test_has_separate_cfg_follows_runtime 测试，验证 Krea-2 Raw 和 Turbo 使用不同 has_separate_cfg 值时 adapter 正确设置；同时修改已有测试以适配新函数签名。

```

def test_has_separate_cfg_follows_runtime(self):
    # 验证 has_separate_cfg 由运行时参数决定，而非模型类硬编码
    module = _import_module_with_stub()
    blocks = ["block_0", "block_1"]

    # 模拟 Krea-2 Raw (CFG=True)
    transformer_raw = _make_transformer("Krea2Transformer2DModel")
    transformer_raw.transformer_blocks = blocks
    adapter_raw = module._build_custom_block_adapter(
        transformer_raw, has_separate_cfg=True
    )
    self.assertTrue(adapter_raw.has_separate_cfg)

    # 模拟 Krea-2 Turbo (CFG=False)
    transformer_turbo = _make_transformer("Krea2Transformer2DModel")
    transformer_turbo.transformer_blocks = blocks
    adapter_turbo = module._build_custom_block_adapter(
        transformer_turbo, has_separate_cfg=False
    )
    self.assertFalse(adapter_turbo.has_separate_cfg)

```

评论区精华

PR 获得两位 reviewer (mickqian, zijiexia) 的快速批准，无公开讨论记录。

- 暂无高价值评论线程

风险与影响

- 风险：核心风险在于 `has_separate_cfg` 的传递路径：若 `batch.do_classifier_free_guidance` 在特定调度场景下为空或错误，可能导致缓存行为偏离预期（但 PR 备注指出 `mismatch` 仅禁用缓存，不会损坏输出）。另外，修改了 `_CUSTOM_BLOCK_ADAPTER_SPECS` 的合约（从三元组变为二元组），任何外部依赖该字典的代码（如有）需同步更新。最后，文档中新增的大量表格和配置示例可能存在渲染问题，但无功能风险。
- 影响：用户影响：使用 Krea-2 或 ERNIE-Image 系列（含 Turbo 变体）的用户可直接受益于 Cache-DiT，加速显著（1.4x~2.9x）。系统影响：Cache-DiT 集成层现在支持按批次动态设置 CFG 模式，为未来更多模型族（如 raw/turbo 对）提供了可复用机制。团队影响：降低了维护成本——无需为每个模型变体单独注册 `has_separate_cfg`。
- 风险标记：核心路径变更，CFG 模式传递依赖，字典合约变更

关联脉络

- PR #28266 [diffusion][cache-dit] initial integration: PR 中提及的早期自定义 `BlockAdapter` 路径，本 PR 在其上扩展并修复了硬编码问题。