

PR #32148 完整报告

sgl-project/sglang

[Kernel] Classification cleanup: unify _jit_ naming, drop empty/model groups, add elementwise (RFC #29630)

合并时间: 2026-07-23 13:47

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32148>

执行摘要

- 一句话: 内核分类清理: 统一 jit 命名、移除空 / 模型分组、新增 elementwise 分组
- 推荐动作: 推荐团队成员关注此 PR 了解新的内核分组规范。刷新了对命名和分类一致性重要性的认识。

功能与动机

在 RFC #29630 完成内核统一命名空间迁移后, PR #32072 和 #32128 将 jit_kernel 合并进 kernels。本 PR 进一步收紧分类和命名一致性: 去掉 batch-1 shim 时代遗留的 jit 前缀, 将 model 分组的内核按功能自然归类, 删除不再有用的空分组 spatial, 新增跨切片的 pointwise 运算分组 elementwise, 使 ops 分组从 19 个减少到 18 个, 命名更一致, 便于后续维护。

实现拆解

1. 统一 JIT 命名: 将 8 个模块从 _jit_*.py 重命名为 *.py, 同步更新内部函数名和注册键。
涉及文件: activation/activation.py、layernorm/norm.py、
gemm/dsv3_{fused_a,router}gemm.py、kvcache/set_mla_kv_buffer.py、
quantization/per{tensor_quant_fp8,token_group_quant,token_group_quant_8bit_v2}.py。
2. 删除空分组 spatial: 移除 python/sglang/kernels/ops/spatial/__init__.py 及其注册的两个内核 (get_sm_available, create_greenctx_stream_by_value), 这些内核由 sgl_kernel wheel 直接提供, 无需在 kernels 中注册。
3. 移除 model 分组: 将 inkling 相关的内核按功能重新分布到 communication (inkling_all_reduce, inkling_ar_fused, inkling_ar_scattered_sconv)、attention (inkling_attn_prologue, inkling_rel_proj, inkling_row_scale)、moe (gate_topk_renorm) 分组。
4. 新增 elementwise 分组: 将 attention/add_constant.py (通用 JIT 参考内核) 和 layernorm/elementwise.py (softcap/sigmoid-mul/gated-activation/fused-rmsnorm) 移入新的 elementwise 分组。
5. 移动 hadamard: 将 attention/hadamard.py 移至 quantization/, 作为量化旋转原语。
6. 更新所有 import 引用: 修改了 90+ 个源文件中的 import 语句和文档注释中的模块路径, 确保无 dangling ref。

7. 修复测试：第二个提交更新了 `test_kernels_namespace.py` 中的 GROUPS 守卫，反映分组增减。

关键文件：

- `python/sglang/kernels/ops/spatial/__init__.py`（模块 spatial 分组；类别 infra；类型 deletion；符号 `get_sm_available`, `create_greenctx_stream_by_value`）：删除了整个空分组 spatial，包含两个内核注册（`get_sm_available`, `create_greenctx_stream_by_value`），体现分组精简。
- `python/sglang/srt/models/inkling_common/attn.py`（模块 模型层；类别 source；类型 data-contract）：更新了 inkling 内核的 import 路径，从 `model.inkling` 迁移至 `attention` 分组，并同步修改了文档注释。
- `python/sglang/srt/models/deepseek_v2.py`（模块 模型层；类别 source；类型 data-contract）：去掉了 `jit` 前缀的导入和调用别名，体现了 JIT 命名统一。
- `python/sglang/srt/models/inkling_common/kernels/comm.py`（模块 通信模块；类别 source；类型 data-contract）：更新了所有 inkling 通信内核的 import 路径，从 `model.inkling` 移至 `communication` 分组。
- `python/sglang/srt/models/inkling.py`（模块 模型层；类别 source；类型 data-contract）：更新了编译时 import 的模块路径，反映分组迁移。
- `python/sglang/kernels/ops/kvcache/set_mla_kv_buffer.py`（模块 内核分组；类别 infra；类型 rename-or-move；符号 `jit_set_mla_kv_buffer_module`, `set_mla_kv_buffer_module`）：重命名文件，去掉 `_jit` 前缀，统一命名风格。
- `python/sglang/kernels/ops/gemm/dsv3_router_gemm.py`（模块 内核分组；类别 infra；类型 rename-or-move；符号 `jit_dsv3_router_gemm_module`, `dsv3_router_gemm_module`）：重命名文件，去掉 `_jit` 前缀，统一命名风格。
- `python/sglang/kernels/ops/activation/activation.py`（模块 内核分组；类别 infra；类型 rename-or-move；符号 `jit_activation_module`, `activation_module`）：重命名文件，去掉 `_jit` 前缀，统一命名风格。

关键符号：`set_mla_kv_buffer_module`, `activation_module`, `dsv3_router_gemm_module`, `dsv3_fused_a_gemm_module`, `inkling_all_reduce`, `inkling_ar_fused`, `inkling_ar_scattered_sconv`, `inkling_attn_prologue_verify`, `inkling_attn_prologue_extend`, `inkling_attn_prologue_decode`, `rel_proj_small_t`, `row_compact_bf16`, `get_sm_available`, `create_greenctx_stream_by_value`

关键源码片段

`python/sglang/srt/models/inkling_common/attn.py`

更新了 inkling 内核的 import 路径，从 `model.inkling` 迁移至 `attention` 分组，并同步修改了文档注释。

```
# attn.py (head) — inkling 内核从 model 分组迁移至 attention 分组
# 之前: from sglang.kernels.ops.model.inkling.inkling_rel_proj import rel_proj_small_t
# 之前: from sglang.kernels.ops.model.inkling.inkling_row_scale import row_compact_bf16
# 现在: 直接归属 attention 分组, 更直观
```

```

from sglang.kernels.ops.attention.inkling_rel_proj import rel_proj_small_t
from sglang.kernels.ops.attention.inkling_row_scale import row_compact_bf16

# 在 docstring 中更新模块路径说明
def _fused_attn_prologue_verify(self, q, k, v, forward_batch, log_scaling_tau=None):
    """ ...
    (kernels/ops/attention/inkling_attn_prologue.py); # 路径从 model.inkling 改为 attention
    ... """
    from sglang.kernels.ops.attention.inkling_attn_prologue import (
        inkling_attn_prologue_verify,
    )
    # ...

```

python/sglang/srt/models/deepseek_v2.py

去掉了 jit 前缀的导入和调用别名，体现了 JIT 命名统一。

```

# deepseek_v2.py (head) — 统一的 dsv3_router_gemm 导入与使用
import torch

# 统一命名：不再使用 _jit_ 前缀，as 别名与原始名一致
if _is_cuda:
    from sglang.kernels.ops.gemm.dsv3_router_gemm import (
        dsv3_router_gemm as dsv3_router_gemm,
    )

def forward(self, hidden_states, forward_batch):
    # ...
    if (
        _is_cuda
        and hidden_states.shape[0] <= max_router_gemm_tokens
        and hidden_states.shape[1] % 1024 == 0
        and (self.weight.shape[0] == 256 or self.weight.shape[0] == 384)
        and _device_sm >= 90
    ):
        # 直接使用 dsv3_router_gemm，不再包装 _jit_dsv3_router_gemm
        logits = dsv3_router_gemm(
            hidden_states, self.weight, out_dtype=torch.float32
        )
    # ...

```

评论区精华

无 review 讨论，PR 由作者直接合并。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险是大量重命名和 import 更新可能导致遗漏或错误。但 PR 通过了 py_compile + pre-commit (isort/ruff/black) 以及 registry-validation 和 no-registered-tests-in-package 检查，且所有 18 个分组都能加载。由于不涉及运行时逻辑变更，功能回归风险低。但若其他未覆盖的分支（如 benchmark 脚本、第三方扩展）使用了旧路径，可能产生构建失败。
- 影响：影响范围：对开发者而言，内核分类和命名更加一致，降低了认知负担；对依赖 kernels 包的外部用户，若直接引用了旧的模块路径（如 `sglang.kernels.ops.model.inkling.xxx`），需要相应更新。对外部用户无明显影响，因为公开 API 保持不变。团队维护成本降低。
- 风险标记：大量文件变动，外部引用需更新，未覆盖 benchmark/thirdparty

关联脉络

- PR #32072 [Kernel] RFC #29630 finale: retire `sglang.jit_kernel` into `sglang.kernels`: 本 PR 是 RFC #29630 的后续清理，#32072 完成了 `jit_kernel` 的迁移，本 PR 在此基础上进行命名统一和分组优化。
- PR #32128 [Kernel] Reclassify kernel tests by ops group + move helpers out of the package (RFC #29630): #32128 对内核测试按 ops 组重新分类，与本次分组清理互补，共同收尾 RFC #29630。