PR #26859 完整报告

sgl-project/sglang

FrozenKVMTPVerifyInput: add _draft_preprocess_idle call for when all requests in the verify batch finish in the same iteration

合并时间：2026-06-05 12:47

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26859

执行摘要

- 一句话：修复 Frozen-KV MTP 验证批量全部完成时缺少 merge_batch 的崩溃
- 推荐动作：建议精读此 PR，了解 Frozen-KV MTP 中 verify 和 draft 输入的生命周期管理。其中的空闲输入处理模式（create_idle_input）在其他推测解码实现中也有类似应用，值得参考。对于使用 Frozen-KV MTP 的团队，建议尽快合并。

功能与动机

修复 Issue #27007 中报告的服务器崩溃问题：当使用 Frozen-KV MTP 推测解码且批量中所有请求同时完成（如工具调用或停止序列）时，调度器在下一轮合并批次时因 FrozenKVMTPVerifyInput 缺少 merge_batch 属性而抛出 AttributeError，导致服务中断。

实现拆解

1. 在 frozen_kv_mtp_worker.py 中新增 _draft_preprocess_idle 方法，将 batch.spec_info 替换为一个空闲的 FrozenKVMTPDraftInput（通过 create_idle_input 工厂方法创建），确保 merge_batch 和 filter_batch 在下一轮调度器操作中可用。
2. 在 forward_batch_generation 方法的 else 分支（对应 draft_extend_input.input_ids.shape[0] == 0 且 enable_dp_attention 为 False 的情况）调用 _draft_preprocess_idle，取代原来的仅跳过逻辑。
3. 新增单元测试文件 test_frozen_kv_mtp_all_reqs_finish_in_verify.py，使用 Mock 模拟 FrozenKVMTPWorker 的 draft 和 verify 步骤，验证当验证输出返回空 input_ids 时，forward_batch_generation 正确安装空闲 draft 输入，并且后续的 merge_batch 操作不抛出异常。

关键文件：

- python/sglang/srt/speculative/frozen_kv_mtp_worker.py（模块 推测解码；类别 source；类型 core-logic；符号 _draft_preprocess_idle, forward_batch_generation）：核心修复文件，新增 _draft_preprocess_idle 方法并在 forward_batch_generation 的 else 分支调用，是解决崩溃的关键变更。
- test/registered/unit/spec/test_frozen_kv_mtp_all_reqs_finish_in_verify.py（模块 单元测试；类别 test；类型 test-coverage；符号 TestFrozenKVMTPWorker, _make_worker, _make_decode_batch, _forward_generation）：新增的单元测试文件，覆盖了修复后的边

界条件，确保所有请求完成时能正确安装空闲 draft 输入且后续 merge_batch 不崩溃。

关键符号：_draft_preprocess_idle, forward_batch_generation

关键源码片段

python/sglang/srt/speculative/frozen_kv_mtp_worker.py

核心修复文件，新增 _draft_preprocess_idle 方法并在 forward_batch_generation 的 else 分支调用，是解决崩溃的关键变更。

```
# frozen_kv_mtp_worker.py (partial)

def _draft_preprocess_idle(self, batch: ScheduleBatch) -> None:
    # 当验证批量中所有请求在同一轮完成时，将 batch.spec_info 替换为
    # 一个空闲的 FrozenKVMTPDraftInput，使得下一轮调度器的
    # merge_batch / filter_batch 能正确处理（避免 AttributeError）。
    batch.spec_info = FrozenKVMTPDraftInput.create_idle_input(
        device=self.device,
        hidden_size=self._recurrent_hidden_size,
        dtype=self.model_config.dtype,
        topk=self.topk,
        capture_hidden_mode=CaptureHiddenMode.LAST,
    )

def forward_batch_generation(self, batch: ScheduleBatch) -> GenerationBatchResult:
    # ... ( 前面的 verify 逻辑 )
    with (
        self.draft_tp_context(self.draft_model_runner.tp_group),
        speculative_moe_backend_context(),
        speculative_moe_a2a_backend_context(),
    ):
        draft_extend_input = verify_output.draft_extend_input
        if (
            self.server_args.enable_dp_attention
            or draft_extend_input.input_ids.shape[0] > 0
        ):
            # 正常情况：还有未完成的请求，继续 draft extend
            batch.spec_info = draft_extend_input
            self.forward_draft_extend_after_decode(batch)
        else:
            # 所有请求完成且 dp_attention 未强制 extend：安装空闲 draft 输入
            self._draft_preprocess_idle(batch)
    # ... ( 返回结果 )
```

test/registered/unit/spec/test_frozen_kv_mtp_all_reqs_finish_in_verify.py

新增的单元测试文件，覆盖了修复后的边界条件，确保所有请求完成时能正确安装空闲 draft 输入且后续 merge_batch 不崩溃。

```
# test_frozen_kv_mtp_all_reqs_finish_in_verify.py ( 核心测试类 )
```

```
class TestFrozenKVMTPWorker(CustomTestCase):
    def _make_worker(self):
        worker = FrozenKVMTPWorker.__new__(FrozenKVMTPWorker)
        # 配置 worker 为 CPU 模式，简化 Mock
        worker.device = torch.device("cpu")
        worker.topk = TOPK
        worker.model_config = SimpleNamespace(dtype=torch.float32)
        worker.server_args = SimpleNamespace(enable_dp_attention=False)
        worker._model_runner = SimpleNamespace(
            tp_group=None, model=SimpleNamespace(backbone_hidden_size=HIDDEN_SIZE)
        )
        worker.draft_tp_context = lambda _: nullcontext()

        stale_verify = FrozenKVMTPVerifyInput.__new__(FrozenKVMTPVerifyInput)
        worker.draft = Mock(return_value=stale_verify)
        # 模拟验证输出：draft_extend_input.input_ids 为空（所有请求完成）
        worker.verify = Mock(
            return_value=_FakeVerifyOutput(
                draft_extend_input=SimpleNamespace(
                    input_ids=torch.empty((0,), dtype=torch.int64)
                ),
                logits_output=SimpleNamespace(),
                accept_tokens=torch.empty((0,), dtype=torch.int64),
                num_correct_drafts_per_req_cpu=[0, 0],
                can_run_cuda_graph=False,
            )
        )
        worker.forward_draft_extend_after_decode = Mock()
        return worker, stale_verify

    def test_forward_generation_installs_idle_draft_when_verify_finishes_all_reqs(self):
        worker, stale_verify = self._make_worker()
        batch = self._make_decode_batch()
        result = self._forward_generation(worker, batch)
        # 验证：batch.spec_info 不再是 stale 的 FrozenKVMTPVerifyInput，而是
        FrozenKVMTPDraftInput
        self.assertIsNot(batch.spec_info, stale_verify)
        self.assertIsInstance(batch.spec_info, FrozenKVMTPDraftInput)
        # 验证空闲输入的 tensor 形状为空
        self.assertEqual(batch.spec_info.topk_index.shape, (0, TOPK))
        self.assertEqual(batch.spec_info.hidden_states.shape, (0, HIDDEN_SIZE))
        # 验证后续 merge_batch 不抛出异常
        example_prefill = _make_prefill_draft_input()
        batch.spec_info.merge_batch(example_prefill) # 应正常通过
```

评论区精华

- 审核者 kpham-sgl 要求提供可复现脚本，贡献者 akelch11 给出了端到端复现命令和单元级复现脚本。

- kpmam-sgl 最初对单元测试的必要性存疑，但在重新考虑后认可，并要求重命名测试文件使其更具描述性。
- gemini-code-assist[bot] 在自动代码审查中指出注释中 'simulatenously' 的拼写错误，建议改为 'simultaneously'。
- 可复现脚本确认 (question): 提供了两种复现方式：GPU 端到端和 CPU 单元测试，确认问题可复现且修复有效。
- 单元测试文件命名 (design): 文件被重命名，测试内容保留，获得最终批准。
- 注释拼写错误 (style): 贡献者接受了建议，在后续版本中修复了拼写。

风险与影响

- 风险：低风险。仅修改了特定分支下的行为，原有逻辑不受影响。新增的 _idle_input 创建与 _run_assistant_seed_step 中的空闲处理模式一致，共享 create_idle_input 工厂方法，确保一致性。单元测试覆盖了核心路径，但未包含端到端集成测试。若 create_idle_input 的 shape 与后续预期不符可能导致问题，但已验证与现有用法一致。
- 影响：直接影响使用 Frozen-KV MTP 推测解码的用户，特别是当模型频繁触发停止序列或工具调用时，避免了服务器崩溃。不影响其他推测解码模式或其他功能。新增的单元测试可在 CPU 上运行，无需 GPU，降低 CI 成本。修复范围明确，仅涉及一个边界条件。
- 风险标记：边缘条件修复，单元测试覆盖，无端到端测试

关联脉络

- PR #27007 [Bug] FrozenKVMTPVerifyInput missing merge_batch — server crashes on tool calls / stop sequences: 本 PR 直接修复该 Issue 报告的问题。
- PR #27300 fix(spec): complete CustomSpecAlgo duck-typing interface and guard against drift: 同为推测解码模块的修复，展示了类似的接口一致性维护模式。
- PR #27193 Replace skip_attn_backend_init with a batch-carried attention plan marker (+ staleness re-plan): 涉及推测解码中 batch spec_info 的生命周期管理，与本 PR 有概念关联。
- PR #27316 fix(attn): delegate init_mha_chunk_metadata in HybridLinearAttnBackend: 另一个注意力后端修复，与本 PR 同属边界条件修复类别。