PR #34560 完整报告

sgl-project/sglang

[Fix] Fix Qwen3.5 MTP startup with HiCache

合并时间：2026-08-15 01:55

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34560

执行摘要

- 一句话：同步 Qwen3.5 MTP 深度到 text_config，修复 HiCache 启动崩溃
- 推荐动作：值得精读。这是一个小而聚焦的 bugfix 范本：根因定位精确（嵌套 HF config 与 SGLang 派生属性之间的数据契约错位）、修复最小（一行配置同步）、并顺势加固了 review 中发现的 fallback 路径缺陷。关注两个设计点：一是 ModelConfig 从 hf_text_config 派生属性与 _config_draft_model 归一化只写父 config 的隐含契约，同类模型（如未来的 conditional-generation 架构）都可能踩中；二是 HiCache sidecar 路径对池类型的防御性解包，体现了“主路径修复 + fallback 兜底”的双层修正思路。

功能与动机

PR body 明确指出：Qwen3.5 conditional-generation checkpoints store language-model attributes in the nested text_config；MTP draft remapping 时只在父级 Hugging Face config 上设置 num_nextn_predict_layers = 1，而 ModelConfig.num_nextn_predict_layers 派生自 hf_text_config，导致 draft depth 仍为 None，HiCache 下 draft cache 被误判为 sidecar，调度器初始化在 build_full_draft_pools 中访问 pool.layer_num 时崩溃。该回归由 #30393 的 packed-versus-sidecar HiCache draft 路由暴露，原始 Qwen3.5 支持 #18489 只做了一半归一化，本 PR 补全另一半。

实现拆解

1. 配置数据契约修复：在 python/sglang/srt/configs/model_config.py 的 _config_draft_model 方法中，Qwen3.5/InternS2 系列分支（Qwen3_5ForConditionalGeneration、Qwen3_5MoeForConditionalGeneration、Qwen3_5ForCausalLM、Qwen3_5MoeForCausalLM、InternS2PreviewForConditionalGeneration、InternS2MobiusForConditionalGeneration）内，原本只设置 self.hf_config.num_nextn_predict_layers = 1 和 self.hf_config.architectures[0] = "Qwen3_5ForCausalLMMTP"，现新增 self.hf_text_config.num_nextn_predict_layers = 1，使 ModelConfig.num_nextn_predict_layers 这一派生属性也能正确读取到 MTP 深度，保证 HiCache 走 packed draft 池而非误判 sidecar。
2. sidecar fallback 路径加固：在 python/sglang/srt/mem_cache/hybrid_cache/hybrid_pool_assembler.py 的 build_full_draft_pools 中，访问 pool.layer_num 前先判断 isinstance(pool, HybridLinearKVPool)，若是则解包到 pool.full_kv_pool。因为

HybridLinearKVPool 本身没有 layer_num 属性，其唯一注意力层保存在子池中；这样即使 draft 被降级为 sidecar fallback，也不会再触发 AttributeError。该改动来源于 1e4ves 的 review 意见，并在后续 commit 中合入，Co-authored-by 也署名了 1e4ves。

3. 单元回归测试：在 test/registered/unit/configs/test_model_config.py 新增 TestDraftModelConfig.test_qwen35_mtp_depth_is_synced_to_text_config，用 object.__new__ 构造 ModelConfig，验证 _config_draft_model 后 hf_text_config.num_nextn_predict_layers == 1；在 test/registered/unit/mem_cache/test_hybrid_pool_assembler.py 新增 TestDraftSidecarPoolDispatch.test_full_builder_unwraps_empty_hybrid_linear_pool，用 layer_num=0 的空 HybridLinearKVPool 验证解包后返回空 sidecar 列表。
4. E2E 测试增强：test/registered/hicache/test_qwen35_hicache.py 的 Qwen3.5 HiCache 启动测试补上 NEXTN 相关参数（--speculative-algorithm NEXTN、--speculative-num-steps 3、--speculative-eagle-topk 1、--speculative-num-draft-tokens 4），使既有 E2E 用例开始覆盖 NEXTN + HiCache 组合场景，避免回归再次漏检。

关键文件：

- python/sglang/srt/configs/model_config.py（模块 模型配置；类别 source；类型 data-contract；符号 _config_draft_model）：核心修复文件：在 _config_draft_model 中为 Qwen3.5/InternS2 系列补上 hf_text_config.num_nextn_predict_layers = 1，使派生属性 ModelConfig.num_nextn_predict_layers 不再为 None，彻底修复 HiCache draft 误判为 sidecar 的根因。
- python/sglang/srt/mem_cache/hybrid_cache/hybrid_pool_assembler.py（模块 池装配；类别 source；类型 core-logic；符号 build_full_draft_pools）：Sidecar fallback 路径加固：build_full_draft_pools 在访问 layer_num 前解包 HybridLinearKVPool 到其 full_kv_pool，避免误判为 sidecar 时同样触发 AttributeError，由 1e4ves 提出并合入。
- test/registered/unit/configs/test_model_config.py（模块 模型配置；类别 test；类型 test-coverage；符号 TestDraftModelConfig, test_qwen35_mtp_depth_is_synced_to_text_config）：新增 TestDraftModelConfig.test_qwen35_mtp_depth_is_synced_to_text_config，直接验证 hf_text_config.num_nextn_predict_layers 被同步，防止根因再次回归。
- test/registered/unit/mem_cache/test_hybrid_pool_assembler.py（模块 池装配；类别 test；类型 test-coverage；符号 TestDraftSidecarPoolDispatch, test_full_builder_unwraps_empty_hybrid_linear_pool）：新增 TestDraftSidecarPoolDispatch.test_full_builder_unwraps_empty_hybrid_linear_pool，覆盖 HybridLinearKVPool 解包分支，保证 sidecar fallback 不再崩溃。
- test/registered/hicache/test_qwen35_hicache.py（模块 HiCache；类别 test；类型 test-coverage）：E2E 测试补上 NEXTN 启动参数，让 HiCache 集成测试真正覆盖 Qwen3.5 + NEXTN 组合，防止启动崩溃在集成层面再次漏检。

关键符号：_config_draft_model, build_full_draft_pools, test_qwen35_mtp_depth_is_synced_to_text_config, test_full_builder_unwraps_empty_hybrid_linear_pool

关键源码片段

python/sglang/srt/configs/model_config.py

核心修复文件：在 _config_draft_model 中为 Qwen3.5/InternS2 系列补上 hf_text_config.num_nextn_predict_layers = 1，使派生属性 ModelConfig.num_nextn_predict_layers 不再为 None，彻底修复 HiCache draft 误判为 sidecar 的根因。

```
# python/sglang/srt/configs/model_config.py
# _config_draft_model 中的 Qwen3.5/InternS2 分支。
# 关键点：Qwen3.5 conditional-generation checkpoint 把语言模型属性放在
# 嵌套的 text_config 里，而 ModelConfig.num_nextn_predict_layers 是从
# hf_text_config 派生的，所以必须同时写父 config 和 text_config，
# 否则 HiCache 会拿不到 MTP 深度，把 draft 池误判为 sidecar 而在启动时崩溃。
if is_draft_model and self.hf_config.architectures[0] in [
    "Qwen3_5ForConditionalGeneration",
    "Qwen3_5MoeForConditionalGeneration",
    "Qwen3_5ForCausalLM",
    "Qwen3_5MoeForCausalLM",
    "InternS2PreviewForConditionalGeneration",
    "InternS2MobiusForConditionalGeneration",
]:
    if (
        self.hf_config.architectures[0]
        == "InternS2MobiusForConditionalGeneration"
    ):
        # InternS2Mobius 的目标模型拥有 2,560 个专家（四个共享物理 bank），
        # 而其内嵌 MTP 层是普通 Qwen3.5 MoE 层，专家数更小，
        # 需要先用 checkpoint 声明的 mtp_num_experts 覆盖 text_config。
        self.hf_text_config.model_type = "qwen3_5_moe_text"
        self.hf_text_config.num_experts = self.hf_text_config.mtp_num_experts
        self.hf_text_config.num_experts_per_tok = (
            self.hf_text_config.mtp_num_experts_per_tok
        )
    # 归一化：把架构改写为 MTP 变体，并在两个 config 上同时写入 draft 深度。
    self.hf_config.architectures[0] = "Qwen3_5ForCausalLMMTP"
    self.hf_config.num_nextn_predict_layers = 1
    self.hf_text_config.num_nextn_predict_layers = 1
```

python/sglang/srt/mem_cache/hybrid_cache/hybrid_pool_assembler.py

Sidecar fallback 路径加固：build_full_draft_pools 在访问 layer_num 前解包 HybridLinearKVPool 到其 full_kv_pool，避免误判为 sidecar 时同样触发 AttributeError，由 1e4ves 提出并合入。

```
# python/sglang/srt/mem_cache/hybrid_cache/hybrid_pool_assembler.py
# 构建 HiCache draft sidecar 池的入口。
# 注意：draft 池可能是 HybridLinearKVPool，它本身没有 layer_num 属性，
# 唯一的注意力层保存在 full_kv_pool 子池里；若不先解包，
```

```
# 任何走 sidecar fallback 的 draft（例如 MTP 深度尚未归一化时）
# 都会在访问 pool.layer_num 时崩溃。这里做防御性解包。
def build_full_draft_pools(
    *,
    draft_kv_pool: Any,
    tree_cache: Any,
    server_args: ServerArgs,
) -> tuple[list[SidecarPoolSpec], list[PoolEntry]]:
    """Build draft KV/DSA sidecars whose indices follow target full KV."""
    from sglang.srt.mem_cache.memory_pool import (
        DSATokenToKVPool,
        HybridLinearKVPool,
    )

    pool = draft_kv_pool
    if isinstance(pool, HybridLinearKVPool):
        # Hybrid draft runners keep their sole attention layer in this sub-pool.
        pool = pool.full_kv_pool
    if pool.layer_num == 0:
        return [], []

    controller = tree_cache.cache_controller
    host_pool_group = controller.mem_pool_host
    # ... 后续按解包后的 pool 构建 host 池与 sidecar spec
```

test/registered/unit/configs/test_model_config.py

新增 TestDraftModelConfig.test_qwen35_mtp_depth_is_synced_to_text_config，直接验证 hf_text_config.num_nextn_predict_layers 被同步，防止根因再次回归。

```
# test/registered/unit/configs/test_model_config.py
# 回归测试：验证 Qwen3.5 系列 draft 模型归一化时，
# MTP 深度不仅写到父 hf_config，也同步到 hf_text_config。
class TestDraftModelConfig(CustomTestCase):
    def test_qwen35_mtp_depth_is_synced_to_text_config(self):
        # 用 object.__new__ 跳过 __init__，只构造测试所需的属性。
        config = object.__new__(ModelConfig)
        config.is_draft_model = True
        config.speculative_algorithm = "EAGLE"
        config.hf_config = SimpleNamespace(
            architectures=["Qwen3_5MoeForConditionalGeneration"]
        )
        config.hf_text_config = SimpleNamespace()

        config._config_draft_model()

        # 归一化结果：架构改写为 MTP 变体，
        # 父 config 与 text_config 的 draft 深度都必须为 1。
        self.assertEqual(config.hf_config.architectures, ["Qwen3_5ForCausalLMMTP"])
        self.assertEqual(config.hf_config.num_nextn_predict_layers, 1)
```

```
self.assertEqual(config.hf_text_config.num_nextn_predict_layers, 1)
```

评论区精华

评论区最有价值的交锋来自 1e4ves：他说 "thanks! But sidecar should not fail as a fallback path, so there are also bugs here. let me fix it." 即指出：即使 draft 被误判为 sidecar，fallback 路径也不应该崩溃，因此还存在独立缺陷。DarkraiHL 随后回复 "Thanks for catching the fallback-path issue. I've incorporated your HybridLinearKVPool unwrapping fix, added focused regression coverage."，将解包修复与对应回归测试并入本 PR。这表明主修复解决的是配置归一化（root cause），而 fallback 加固解决的是防御性健壮性（same crash, different trigger）。

- Sidecar fallback 路径本身也存在崩溃缺陷 (design): DarkraiHL 采纳 1e4ves 的 HybridLinearKVPool 解包方案，在 build_full_draft_pools 中先解包到 full_kv_pool 再访问 layer_num，并补充对应回归测试与 co-author 署名。
- 根因确认：Qwen3.5 嵌套 text_config 导致 MTP 深度未同步 (correctness): 通过新增 hf_text_config.num_nextn_predict_layers = 1 完成归一化，并在单测中断言该属性值，根因闭环确认。

风险与影响

- 风险：
  1. model_config.py 的新增行依赖 hf_text_config 存在；但该分支内此前已访问 self.hf_text_config.model_type（InternS2Mobius 分支），因此不会引入新的 None 解引用路径，风险可控。
  2. hybrid_pool_assembler.py 假设 HybridLinearKVPool.full_kv_pool 一定存在且代表唯一注意力层；若未来出现多子池的 HybridLinearKVPool，单一解包可能覆盖不全。当前测试仅覆盖 layer_num=0 的空池场景，对非空池的 sidecar 构建路径缺少断言。
  3. 修复会让部分此前崩溃的 Qwen3.5+HiCache 启动变为可运行，但该组合首次在 E2E 中启用 NEXTN，存在超出本次启动阶段的潜在运行时问题（如 draft 缓存读写），需要后续观测。
  4. 多次 merge main（8 个 commit 中 6 次为 merge）可能引入无关变更，但最终 patch 只有 56 行增、3 行删，范围可控。 - 影响：影响用户：Qwen3.5-397B-A17B-FP8 等 Qwen3.5/InternS2 系列模型在 NEXTN + HiCache 组合下从完全无法启动（server 永不健康）变为可正常 serving；同一修复也被外部 SemiAnalysisAI/InferenceX 直接应用到其 MI355X 基准配方，说明对 ROCm/AMD 平台用户同样有效。影响系统：仅启动期配置归一化与 HiCache 池装配逻辑，不触碰推理热路径，无性能影响。影响团队：为 HiCache draft 路由补充了回归防线，后续修改 packed/sidecar 判定时多了一层测试约束。 - 风险标记：调度器初始化路径，嵌套配置契约依赖，fallback 路径加固，新组合场景首次 E2E 覆盖

关联脉络

- PR #30393 Introduce packed-versus-sidecar HiCache draft routing: PR body 明确指出该回归就是 30393 引入的 packed/sidecar 路由导致的：MTP 深度为 None 时 draft 缓存被误判为 sidecar，从而暴露 pool.layer_num 崩溃。
- PR #18489 Original Qwen3.5 MTP support: 原 Qwen3.5 支持只在父级 HF config 上归一化 MTP 深度，本 PR 补全其对 hf_text_config 的同步，属于同一功能的延续修复。
- PR #33810 GDN chunked-extend padding fix: 关联 Issue 2582 的 body 提到，TP2/EP2 MTP 场景依赖该 GDN 修复；它与本 PR 共同构成 Qwen3.5 + MTP + HiCache 组合可用的前提条件。