PR #34655 完整报告

sgl-project/sglang

[diffusion] feat: track MiniMax-H3 in the nightly diffusion benchmark

合并时间：2026-08-13 15:12

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34655

执行摘要

- 一句话：MiniMax-H3 纳入 diffusion nightly 基准并支持定制请求参数
- 推荐动作：值得快速精读，重点看 _build_sglang_payload 的扩展方式：用 null 表示「删除键」而不是「发送 null」，语义清晰且对旧用例完全向后兼容。这种「先扩展纯函数、再注册用例」的提交拆分方式也适合 CI/benchmark 类改动参考。

功能与动机

PR body 指出：MiniMax-H3 是仓库中唯一由 SGLang 服务的联合视频 + 音频模型，且没有任何 nightly 覆盖，“regressions in it stay invisible until someone runs it by hand”。模型本身又拒绝显式 num_frames：“num_frames is not supported: MiniMax H3 derives the temporal shape from target.duration_seconds”，因此必须先扩展 harness 的请求组装逻辑，才能注册用例。

实现拆解

1. 运行器扩展：在 scripts/ci/utils/diffusion/run_comparison.py 的 _build_sglang_payload 末尾新增 sglang_request_extra 处理循环：值为 None 时调用 payload.pop(key, None) 删除键，否则写入或覆盖 payload[key]。由于现有用例都没有设置该字段，case.get("sglang_request_extra") or {} 保证所有其他 payload 保持字节级不变。
2. 用例注册：在 scripts/ci/utils/diffusion/comparison_configs.json 新增 minimax_h3_t2va_5s，模型为 MiniMaxAI/MiniMax-H3，任务 text-to-video，5 秒 1344x768 档位，通过 sglang_request_extra 写入 task、target、flow_shift、audio_flow_shift，并把 num_frames、fps 置 null 以删除公共字段。
3. 拓扑与编译配置：serve_args 使用 cookbook 的 4xH100 拓扑（--tp-size 2 --ulysses-degree 2），并显式设置 --enable-torch-compile false，因为 H3 的 torch.compile 路径会改变数值输出。
4. 验证与 CI：mickqian 在 4xH200 上通过真实 harness 端到端运行（latency_s 约 78 秒、error: null、服务干净退出），同时验证了 null 删除路径；PR Test 与 PR Test Extra 两条 CI 均通过。

关键文件：

- scripts/ci/utils/diffusion/comparison_configs.json（模块 评测配置；类别 infra；类型 configuration）：注册了 MiniMax-H3 的 nightly 对比用例，并通过 sglang_request_extra 定制模型特定请求参数，是本 PR 的核心数据配置。
- scripts/ci/utils/diffusion/run_comparison.py（模块 评测工具；类别 infra；类型 core-logic；符号 _build_sglang_payload）：为 _build_sglang_payload 增加 sglang_request_extra 处理逻辑，是让 MiniMax-H3 用例可运行的关键 harness 改动。

关键符号：_build_sglang_payload

关键源码片段

scripts/ci/utils/diffusion/comparison_configs.json

注册了 MiniMax-H3 的 nightly 对比用例，并通过 sglang_request_extra 定制模型特定请求参数，是本 PR 的核心数据配置。

```
{
  "id": "minimax_h3_t2va_5s",
  "model": "MiniMaxAI/MiniMax-H3",
  "task": "text-to-video",
  "prompt": "At night, while their owner sleeps in a bedroom, three cats march in loudly playing
  tiny brass instruments, then abruptly file out.",
  "width": 1344,
  "height": 768,
  "num_frames": 124,       // 公共字段，稍后会被 sglang_request_extra 中的 null 删除
  "fps": 24,               // 公共字段，同样会被删除
  "num_inference_steps": 50,
  "seed": 1101,
  "num_gpus": 4,
  "sglang_request_extra": {
    "task": "t2va",
    "conditions": [],
    "target": {"short_edge": 768, "aspect_ratio": "16:9", "duration_seconds": 5.0},
    "flow_shift": 12.0,
    "audio_flow_shift": 3.0,
    "num_frames": null,    // null 值让 harness 删除该 key，而不是发送 null
    "fps": null
  },
  "frameworks": {
    "sglang": {
      "serve_args": "--warmup-mode server --model-variant fl2va --tp-size 2 --ulysses-degree 2 --
      performance-mode speed --enable-torch-compile false",
      "extra_env": {}
    }
  }
}
```

scripts/ci/utils/diffusion/run_comparison.py

为 _build_sglang_payload 增加 sglang_request_extra 处理逻辑，是让 MiniMax-H3 用例可运行的关键 harness 改动。

```
def _build_sglang_payload(case: dict) -> dict:
    # 前段逻辑：从公共白名单字段（size、num_inference_steps 等）复制到 payload
    payload = {}

    # 本次新增的模型特定扩展：MiniMax-H3 会从 target 推导时间维度，并且拒绝
    # 显式 num_frames（报错：num_frames is not supported ...）。
    # 因此用例需要既能新增键（task、target、flow_shift 等），也要删除公共键；
    # 约定 null 表示删除该键，避免向服务端发送 null 值。
    for key, value in (case.get('sglang_request_extra') or {}).items():
        if value is None:
            payload.pop(key, None)
        else:
            payload[key] = value

    return payload
```

评论区精华

没有正式的 review 评论，但合并者在 issue 评论中补充了关键验证信息：

- 使用真实 harness 入口点（非手写请求）验证，--dry-run 正确解析出 sglang serve --model-path MiniMaxAI/MiniMax-H3 ... --tp-size 2 --ulysses-degree 2 --performance-mode speed --enable-torch-compile false。
- 真实运行结果：latency_s: 78.172, error: null，服务端正常关闭。
- 该运行实际覆盖了 sglang_request_extra 的 null 移除逻辑：H3 拒绝显式 num_frames，请求成功完成正好证明 key-drop 机制在真实请求中生效。
- 作者说明 devbox 安装包含未合并的 #34385（lingbot-only）分支，但 H3 不触碰 lingbot 代码路径，因此结果可代表 main。
- 端到端验证与分支依赖说明 (other): latency_s 78.172、error null、服务干净关闭，null 删除路径被真实请求覆盖验证通过。

风险与影响

- 风险：
  - 变更仅影响 diffusion nightly 评测工具与配置，不涉及运行时逻辑；sglang_request_extra 默认不存在，所有旧用例的 payload 保持字节级一致。
  - 新用例依赖 H3 对 target 的解析与 null 删除语义；若未来公共 schema 增加新字段且 H3 行为不变，需要同步维护 sglang_request_extra。
  - 用例在 H200 实测约 6.5 分钟（含启动与模型加载），H100 nightly runner 的绝对耗时可能不同，需要关注夜间预算。
  - 显式关闭 torch.compile 是为了数值一致，但意味着该基准只覆盖非编译路径，编译路径的回归不会由此用例暴露。

    - 没有新增单元测试，但作者做了真实 harness 端到端验证；未来改动 _build_sglang_payload 时需留意 null 删除语义。
- 影响：
    - 对用户与线上服务零影响，纯 CI/benchmark 变更。
    - MiniMax-H3 从无回归覆盖变为有 nightly 性能与内存基线，后续 image/video 相关回归会更早暴露。
    - sglang_request_extra 成为一个可复用的模型特定请求参数扩展点，未来音频 / 视频联合模型可直接复用。
    - nightly 评测总时长增加约 6.5 分钟，在可接受范围内。
    - 风险标记：夜间评测预算增加，模型特定参数扩展，torch.compile 路径未覆盖

关联脉络

- PR #34652 [diffusion] feat: publish an index of nightly comparison runs: 同属 scripts/ci/utils/diffusion 基础设施，本 PR 新增的 MiniMax-H3 用例将成为该 nightly 对比索引中的一条记录。
- PR #34328 [AMD][CI] CI: fix AMD 2-GPU multimodal-gen partition-count abort: 同为 diffusion/multimodal-gen 夜间评测 CI 的配套修复，共同完善 diffusion 回归体系。