PR #35016 完整报告

sgl-project/sglang

[diffusion] test: tighten NVIDIA perf baselines

合并时间：2026-08-16 20:54

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/35016

执行摘要

- 一句话：收紧 NVIDIA 基线 1381 项，恢复 Hunyuan3D 全管线覆盖
- 推荐动作：值得精读，尤其是 PR body 的归因表：它示范了性能基线维护的规范动作——用同一 revision 的连贯测量替换 field-wise 散值、对异常 delta 区分 " 配置差异 " 与 " 真实优化 "、剔除失败测量并用重试替代、以独立递归审计校验零上调。对 diffusion CI 维护者，建议关注后续 flaky 情况：若新基线频繁误报，可对尾部敏感场景引入多轮测量取中位数或小幅回退。对性能工程师，turbo_wan/fsdp/ltx 的归因方法（对比内存驻留量、step 时间、模式差异）可直接复用到其他平台基线维护。

功能与动机

PR body 开宗明义：tighten existing H100 and RTX 5090 diffusion performance baselines using fresh multimodal_gen CI measurements，且目标为 lower 1,381 existing numeric thresholds without any increases。深层动机有三：其一，旧基线来自 2026 年 1-6 月的陈旧快照（H100 引用 PR 28127 的 run、5090 引用 PR 29791 的 run），测量环境是 " 无 warmup + 默认 offload 全部视频组件 "，与当前默认路径（request shape warmup、小模型组件驻留、FSDP 分片驻留 GPU）差异巨大，基线失去检测灵敏度；其二，PR #28781 曾把 hunyuan3d_shape_gen 改为 shape-only（注入 paint_enable=false），导致 paint/texture 阶段回归长期无法被 CI 发现，本次恢复完整管线；其三，B200 的条目是占位符而非有效阈值（its entries are placeholders rather than active thresholds），保持不动以避免虚假收紧。

实现拆解

1. 采集基准测量：以源码 revision 2ee0d38a8553d186faad9c50893480e3c1d79270 为基准，跑完 H100 1-GPU 5 个 shard、2-GPU 3 个 shard、RTX 5090 与 B200 的服务器测试，并确认该 revision 与 base revision 之间基线文件未变化，保证新旧值可比。失败的 ltx_2_3_hq_pipeline、ltx_2_3_two_stage_ti2v_2gpus、minimax_h3_t2va_2gpu_h100、qwen_image_edit_2511_ti2i 测量被剔除并替换为成功重试的数据。
2. 重写 H100 基线（perf_baselines/h100.json，+1382/-1379）：51 个场景共 1354 项数值按新测量整体下调，覆盖 stages_ms、denoise_step_ms、expected_e2e_ms、expected_avg_denoise_ms、expected_median_denoise_ms 等全部指标，零上调；metadata 更新为记录新来源 revision 与 2026-08-16 时间戳。tolerances（long_term 15%/10%/50%，pr_test 25%/25%/80%）与 sampling 结构保持不变，只动期望值、不改

容差。

3. 收紧 RTX 5090 基线（perf_baselines/5090.json，+29/-29）：3 个场景 27 项下调，集中在 flux_2_klein_base_image_t2i（首步 219.58→95.07 ms，warmup 相关）、wan2_1_t2v_1.3b（DecodingStage 1202.43→495.01 ms、TextEncodingStage 1210.22→1025.11 ms）与 zimage_image_t2i（各 denoise step 小幅下调）。
4. 恢复 Hunyuan3D 完整管线（test_server_common.py，-16 行）：删除 diffusion_server fixture 中针对 custom_validator == "mesh" 动态生成 {"paint_enable": false} 配置并注入 --config 的逻辑，hunyuan3d_shape_gen 重新执行 shape + paint/texture 全流程；基线补回 3 个 paint 阶段字段，E2E 按完整管线样本（run 31941712078：237.25-250.49 s）取 field-wise min。paint postprocess 保留 11.5 s 的宽松值，因一次合法运行达到 20.50 s，80% 容差上限为 20.70 s。
5. 归因与验证：对 4 个最大 E2E delta（turbo_wan2_1、fsdp-inference、ltx_2.3 两阶段、zimage）逐一归因，区分配置差异与真实优化；验证手段包括独立递归审计（1381 降、0 增、3 个 paint 字段恢复）、PerformanceValidator 通过、git diff --check、pre-commit 与 PR CI 全平台复跑。测试与配置配套即本 PR 全部内容，无生产代码变更。

关键文件：

- python/sglang/multimodal_gen/test/server/perf_baselines/h100.json（模块 性能基线；类别 test；类型 test-coverage）：本次变更主体：51 个场景、1354 项阈值按 2026-08-16 实测整体下调，零上调；metadata 更新数据来源；直接决定 diffusion CI 对 H100 性能回归的判定灵敏度。
- python/sglang/multimodal_gen/test/server/perf_baselines/5090.json（模块 性能基线；类别 test；类型 test-coverage）：RTX 5090 平台 3 个场景、27 项阈值收紧，主要反映 warmup 与组件驻留后的阶段耗时下降；是本次除 H100 外的第二个基线文件。
- python/sglang/multimodal_gen/test/server/test_server_common.py（模块 服务测试；类别 test；类型 test-coverage；符号 diffusion_server）：删除 diffusion_server fixture 中对 mesh 场景注入 paint_enable=false 的配置覆盖逻辑，恢复 hunyuan3d_shape_gen 完整 shape + paint/texture 管线测试，是本 PR 唯一的非数据类逻辑变更。

关键符号：diffusion_server

关键源码片段

python/sglang/multimodal_gen/test/server/perf_baselines/h100.json

本次变更主体：51 个场景、1354 项阈值按 2026-08-16 实测整体下调，零上调；metadata 更新数据来源；直接决定 diffusion CI 对 H100 性能回归的判定灵敏度。

```
{
  // metadata：记录数据来源 revision 与更新时间，替换掉旧 PR 引用（PR 28127）
  "metadata": {
    "model": "Diffusion Server",
    "hardware": "CI H100 80GB pool",
    "description": "Reference numbers tightened from NVIDIA multimodal_gen CI at source revision 2ee0d38a8553d186faad9c50893480e3c1d79270.",
    "last_updated": "2026-08-16"
```

```
    },
    // 容差保持不变：long_term（15% e2e / 10% denoise / 50% 其余）夜间用，
    // pr_test（25% / 25% / 80%）PR CI 用；本次只收紧期望值，不动容差
    "tolerances": {
        "long_term": { "e2e": 0.15, "denoise_stage": 0.1, "non_denoise_stage": 0.5, "denoise_step":
        0.25 },
        "pr_test": { "e2e": 0.25, "denoise_stage": 0.25, "non_denoise_stage": 0.8, "denoise_step":
        0.3 }
    },
    "scenarios": {
        // 示例场景 qwen_image_t2i：DenoisingStage 从 12404.23 收紧到 12402.12，
        // 其余阶段与各 denoise step 均来自 2026-08-16 同一批次 CI 测量，零上调
        "qwen_image_t2i": {
            "stages_ms": {
                "TextEncodingStage": 232.03,
                "DenoisingStage": 12402.12,
                "InputValidationStage": 0.04,
                "LatentPreparationStage": 0.18,
                "TimestepPreparationStage": 14.99,
                "DecodingStage": 17.7
            },
            "denoise_step_ms": {
                "0": 196.19,
                // ... 其余 47 个 step 同样逐项按实测收紧
                "20": 248.84
            }
        }
    }
}
```

python/sglang/multimodal_gen/test/server/test_server_common.py

删除 diffusion_server fixture 中对 mesh 场景注入 paint_enable=false 的配置覆盖逻辑，恢复 hunyuan3d_shape_gen 完整 shape + paint/texture 管线测试，是本 PR 唯一的非数据类逻辑变更。

```
    # Strict ports: fail immediately if port is occupied instead of silently
    # picking another one (which causes the test client to connect to the wrong server).
    extra_args += " --strict-ports"

    # 此前此处有一块针对 custom_validator == "mesh" 的逻辑：
    # 动态生成 {"paint_enable": false} 配置并注入 --config，
    # 把 hunyuan3d_shape_gen 变成只跑 shape 的用例。
    # 该覆盖是 PR #28781 引入的，导致 paint/texture 阶段失去 CI 检测，
    # 本 PR 删除它，让完整 shape + paint/texture 管线重新纳入性能基线。
    for arg in server_args.extras:
        extra_args += f" {arg}"

    # 启动 server；wait_deadline 由环境变量控制，
    # diffusion 长尾用例（如 Hunyuan3D 全管线 250 s）依赖此上限
```

```
    wait_deadline = float(os.environ.get("SGLANG_TEST_WAIT_SECS", "1200"))
    logger.info(
        "[server-test] Starting server for test case: %s\n"
        "  Model: %s\n"
        "  Port: %s\n"
        "  Wait deadline: %ss\n"
        "  Extra args: %s",
        case.id,
        server_args.model_path,
        port,
        wait_deadline,
        extra_args,
    )

    manager = ServerManager(
        model=server_args.model_path,
        port=port,
        wait_deadline=wait_deadline,
        extra_args=extra_args,
    )
```

评论区精华

该 PR 没有任何 review 评论（comments 与 review comments 均为 0），核心论证全部集中在 PR body 中，作者以自我审计形式回答了 reviewer 通常会追问的问题：

- 方法论声明："These baseline changes should not be read as isolated model speedups. The baseline file is tightened field-by-field, so historical stage minima and E2E values do not necessarily come from one coherent run."——field-wise 收紧不等同于单次连贯运行的加速，避免把 35x 之类字段 delta 误读为模型级加速。
- 覆盖回退发现：35x 的 Hunyuan3D delta 实际暴露的是 PR #28781 把该 case 改为 shape-only 的覆盖回退，本次删除 paint_enable=false 注入并恢复 3 个 paint-stage 阈值。
- 大 delta 归因：turbo_wan2_1（1 月基线无 warmup + 旧版全量 offload，现值 0.96 s，5090 上为 2.59 s，说明非硬件无关）、fsdp-inference（旧 run 激活 FSDP CPUOffloadPolicy 仅 0.52 GiB/rank 驻留、263 ms/step，现行配置 6.30 GiB/rank、83 ms/step，PR CI 复现 1.058 s）、ltx_2.3（旧 snapshot 模式 + DiT offload vs 现行 original 两阶段 + resident DiT，PR CI 重试 9.04 s 复现 6.64 s）、zimage（1.425 s 系人工放宽值，真实改进约 22-29% 而非 47%）。
- 验证口径：独立递归审计 1381 降 / 0 增，完整 Hunyuan3D 样本通过 PerformanceValidator，并附 PR CI 复现链接。
- 暂无高价值评论线程

风险与影响

- 风险：

1. CI 误报风险（h100.json）：1354 项阈值整体收紧且主要依赖 8 月 16 日单批次测量；LTX 场景首次测量失败、以重试数据替代，PR CI 中 9.04 s vs 6.64 s 的抖动佐证尾延迟噪声，未来硬件 / 驱动抖动可能触发非回归性失败。
2. 配置耦合风险：turbo_wan、fsdp-inference、ltx_2.3 等大 delta 来自默认路径变化（warmup、组件驻留、FSDP 驻留），而非内核本身的加速；任何后续默认参数回调（如驻留策略改回 offload 优先）都会造成大范围基线失效，需配套更新。
3. Hunyuan3D覆盖恢复的不确定性：paint阶段只有2个完整样本（run31941712078），paint postprocess 阈值实际是 80% 容差放宽后的值（20.70 s 上限 vs 11.5 s 期望），统计基础偏弱，后续该阶段回归可能误判。
4. B200 覆盖空洞：B200 仍为占位符且性能检查被禁用，Blackwell 平台的回归防护依然缺失。
5. 跨硬件可比性：turbo_wan 在 H100（0.96 s）与 5090（2.59 s）的巨大差异说明基线对后端选择敏感，跨硬件迁移时不能直接类比。 - 影响：CI 灵敏度：收紧后 diffusion PR 的性能回退更容易被捕获，所有涉及 diffusion 的 PR 都会经由新基线判定；25% PR 容差与 15% long-term 容差保留，灵敏度提升主要来自期望值对齐当前默认路径。Hunyuan3D 回归防护：恢复 shape + paint/texture 完整管线，paint 预处理与纹理生成阶段的回归重新纳入检测，修复了 PR #28781 引入的覆盖缺口。开发体验：新基线不再与过时配置（无 warmup、全量 offload）比较，性能 PR（如 QKNorm+RoPE、VAE decode 提速）更容易获得准确判定。无生产影响：仅测试数据与测试框架变更，对运行时、API、模型输出零影响。 - 风险标记：阈值整体收紧 1381 项，基线源于单批次实测，恢复完整管线覆盖，失败测量以重试替代，B200 基线仍为占位符

关联脉络

- PR #34817 [Diffusion] Speed up MiniMax-H3 VAE decode on 2×H100: 同一份 perf_baselines/h100.json 的维护者，将 VAE decode 大幅提速后的收益固化进 H100 基线，与本 PR 属同一条基线演进线。
- PR #34736 [Diffusion] Unify component residency controls: 组件驻留统一控制是本 PR 中 fsdp-inference、turbo_wan 等大 delta 归因的配置根源，两次变更在默认运行路径上前后衔接。
- PR #34929 [Diffusion] Enable breakable CUDA graphs for LTX-2.3: LTX-2.3 breakable CUDA graphs 上线后默认路径提速，与本 PR 中 ltx_2_3_two_stage_t2v_2gpus 基线收紧直接对应。
- PR #34663 [Diffusion] Refresh docs, retire stale knobs, and fix nightly attribution: 同一 diffusion CI/ 基线体系的持续建设，清理过时 server args 并修复 nightly 归因，与本 PR 的基线刷新互补。