PR #21841 完整报告

sgl-project/sglang

[VLM] Add VLM TP=4 per-commit CI test and improve MMMU eval prompt/parser

合并时间：2026-04-02 11:09

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21841

执行摘要

本 PR 通过新增 VLM TP=4 CI 测试和优化 MMMU 评估 prompt 与解析器，显著提升多模态测试覆盖和评估准确性，将准确率从约 0.34 提升至 0.75-0.78，延迟从 145 秒降低至 28 秒，为 VLM 开发提供更可靠的测试基准。

功能与动机

动机源于填补 4-GPU 测试套件中 VLM 测试的空白，并改进 MMMU 评估的准确性和效率。PR body 明确指出："Add Qwen3.5-27B VLM TP=4 MMMU test to stage-c-test-4-gpu-h100 per-commit CI (previously no VLM tests in any 4-GPU suite)" 和 "Improve MMMU eval prompt... to increase accuracy from ~0.34 to ~0.75-0.78"。这些变更参考了 Kimi-Vendor-Verifier 的 CoT 指令格式，旨在标准化答案提取。

实现拆解

主要改动集中在两个文件：

- python/sglang/test/simple_eval_mmmu_vlm.py：
  - prompt 构建：替换原有简单提示为包含 CoT 指令的格式，例如：
- 解析器优化：在 _parse_multi_choice_response 函数中添加正则表达式匹配，优先查找 "Answer: X" 模式： answer_matches = re.findall(r"[Aa]nswer\\s*:\\s**?*?\\s*\\(?([A-Z])\\)?", response) if answer_matches: candidate = answer_matches[-1] if candidate in all_choices: return candidate
- test/registered/vlm/test_vlm_tp4.py：
  - 新增测试类 TestVLMTP4，配置 Qwen3.5-27B 模型在 TP=4 下运行，设置服务器参数并执行 MMMU 评估。
  - 关键参数包括 --tp-size 4 和准确性阈值验证（MMMU_ACCURACY_THRESHOLD = 0.65）。

评论区精华

gemini-code-assist[bot] 在 review 中提出两点改进：

> "The regex is currently case-sensitive for the answer letter ([A-Z]). While the prompt instructs the model to use a specific format, models may occasionally output lowercase letters (e.g., Answer: a). Making this check case-insensitive would improve the robustness of the parser."

> "These arguments (--mamba-*) are specific to Mamba/SSM architecture models. Since Qwen3.5 is a Transformer-based model, these flags are irrelevant and should be removed to avoid confusion and ensure the configuration is clean."

这些建议聚焦于正确性和设计优化，可能已在合并前采纳，以增强代码健壮性和可维护性。

风险与影响

- 风险：解析器变更可能意外影响其他依赖 MMMU 评估的测试，如果正则表达式未完全覆盖模型输出变体，可能导致解析失败。新增 CI 测试可能增加运行时资源消耗，影响 CI 流水线效率。prompt 指令的普适性有限，可能不适用于所有 VLM 模型。
- 影响：正面影响显著，提升了 VLM 测试的全面性和评估准确性，为团队提供了更可靠的性能基准；负面影响可控，但需监控 CI 稳定性和评估一致性。

关联脉络

从历史 PR 看，本 PR 是 VLM 和多模态测试演进的一部分：

- PR 21767：添加 VLM 相关 CI 测试，扩展量化模型覆盖，与本 PR 共同完善测试基础设施。
- PR 21873：优化评估测试的网络超时设置，与本 PR 的评估逻辑改进相辅相成，提升测试鲁棒性。 这些关联显示团队在持续增强多模态测试的覆盖和可靠性，为未来功能开发奠定基础。