PR #23052 完整报告

sgl-project/sglang

[diffusion] doc: update doc

合并时间：2026-04-17 16:23

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23052

执行摘要

本次 PR 全面更新了 SGLang 扩散模型的文档，涵盖 OpenAI API 示例、环境变量、模型兼容性、attention backends 和平台安装指南，旨在提升文档完整性和用户体验，无代码变更，风险较低。

功能与动机

PR 动机是扩展 SGLang Diffusion 的文档覆盖，以支持新功能如 image-to-video 生成、多平台环境配置和模型更新。根据 review 评论，本次变更“显著扩展了文档”，帮助用户更好地使用扩散功能。

实现拆解

1. API 文档更新：在 docs/diffusion/api/openai_api.md 中新增 image-to-video API 示例，支持 multipart form upload 和 reference URL 两种方式，示例代码如下：
```markdownCreate a video (image-to-video)

For I2V or TI2V models (e.g., Wan2.1 I2V, LTX-2.3 two-stage), pass an input image via multipart form upload or a reference URL.

Curl Example (multipart form upload):

bash curl -sS -X POST "http://localhost:30010/v1/videos" \ -H "Authorization: Bearer sk-proj-1234567890" \ -F "prompt=A cat playing a piano" \ -F "input_reference=@input_image.png" \ -F "size=1280x720"

Curl Example (reference URL):

bash curl -sS -X POST "http://localhost:30010/v1/videos" \ -H "Content-Type: application/json" \ -H "Authorization: Bearer sk-proj-1234567890" \ -d '{ "prompt": "A cat playing a piano", "reference_url": "https://example.com/input_image.png", "size": "1280x720" }' `` 2. **环境变量扩展**：修改docs/diffusion/environment_variables.md，新增运行时变量（如SGLANG_DIFFUSION_TARGET_DEVICE）和平台特定变量（如 ROCm、量化相关），组织为结构化表格。 3. **模型兼容性矩阵**：更新 docs/diffusion/compatibility_matrix.md，添加 Wan2.1 Fun、Helios 系列等新模型，并修正 LTX 模型注释语法。 4. **attention backends 文档**：在 docs/diffusion/performance/attention_backends.md中补充新 backend（如sla_attn）和平台支持（Intel XPU、MUSA），更新支持矩阵。 5. **安装指南补充**：在
```

docs/diffusion/installation.md中添加 Intel XPU 安装步骤，扩展多平台支持。 6. **CLI 参数更新**：在docs/diffusion/api/cli.md中新增--image-path参数说明，用于 image-to-video 和 image-to-image 生成。 7. **索引微调**：对docs/diffusion/index.md` 进行微小调整，确保文档一致性。所有变更均为纯文档更新，无测试或代码配套改动。

评论区精华

review 中仅有一个讨论线程：reviewer 建议修正 docs/diffusion/compatibility_matrix.md 中的语法错误，将“uses”改为“use”以保持主谓一致。该建议被接受并应用，无其他争议或深度技术讨论。

风险与影响

风险主要在于文档准确性，如环境变量描述错误可能导致用户配置失误，但无代码执行、性能或安全风险。影响方面，本次更新全面覆盖扩散功能文档，帮助用户更高效地使用新 API、配置环境和平台，提升用户体验和团队文档质量。

关联脉络

与近期 PR #23028（“[codex] Update diffusion skills”）相关，同为扩散模型文档更新，共同完善扩散文档生态。历史 PR 中多次涉及 diffusion 模块（如 PR 22952、23028），显示团队持续优化扩散功能支持，本 PR 是这一趋势的文档补充部分。