

PR #34497 完整报告

sgl-project/sglang

[diffusion][docs] Update Cosmos3 Edge and distilled cookbook

合并时间: 2026-08-12 10:51

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34497>

执行摘要

- 一句话: 更新 Cosmos3 Edge 与蒸馏版 cookbook 文档
- 推荐动作: 值得快速浏览: 若团队正在部署或计划接入 Cosmos3 Edge / 蒸馏版检查点, 本文档是权威配置参考。技术深度较低 (纯文档), 但 "配置驱动变体识别" 与 "蒸馏版固定 sigma 调度" 两个设计语义值得留意, 后续排查用户误配时可快速定位到文档对应小节。

功能与动机

PR body 明确列出三个目标: document the Cosmos3 Edge and Edge Policy DROID checkpoints; add launch and request recipes for the Super 4-step T2I and I2V checkpoints; explain config-driven variant detection, Edge sampling defaults, and distilled fixed-sigma behavior。该 PR 是 #31590 (Add Cosmos3 Edge and Distilled checkpoints support) 的 follow-up: #31590 已在代码层加入 Edge 与蒸馏变体支持及启动期配置识别逻辑, 但用户缺少相应部署指引, 尤其是蒸馏版 "固定 4 步 sigma、强制 guidance_scale=1.0、不可调 num_inference_steps 与 flow_shift" 这类反直觉行为, 不文档化极易被误配。

实现拆解

1. 扩充模型支持表格: 在 docs/cookbook/diffusion/Cosmos/Cosmos3.mdx 的状态表中新增 4 行——nvidia/Cosmos3-Edge (4B 稠密, 覆盖 T2I/T2V/I2V/V2V/action)、nvidia/Cosmos3-Edge-Policy-DROID、nvidia/Cosmos3-Super-Text2Image-4Step 与 nvidia/Cosmos3-Super-Image2Video-4Step (均标注为固定 4 步蒸馏); 同时把正文中的 Cosmos3OmniDiffusersPipeline 内部类名收敛为 "native Cosmos3 pipeline", 降低文档与实现类名的耦合。
2. 新增 Edge checkpoints 小节: 说明 Edge 是 4B 稠密模型、单卡即可服务, 训练面向 256p/480p; 给出默认配置 (视频 832x480 + guidance_scale=5.0, 图像 640x640 + guidance_scale=7.0) 与 9 种支持尺寸; Edge-Policy-DROID 只需替换模型路径, 同一命令可复用。
3. 新增 Distilled checkpoints 小节: 说明蒸馏版 Super 为 64B 模型、建议多卡 (示例 --num-gpus 4); 强调 SGLang 启动时读取 scheduler/scheduler_config.json 自动识别蒸馏变体, 使用检查点内置固定 4 步 sigma 调度并强制 guidance_scale=1.0, 明确 "不要调 num_inference_steps 或 flow_shift"。

4. 更新请求示例：为蒸馏版 T2I 增加 curl 示例（省略调度器控制字段、仅传 `guidance_scale=1.0`）；为蒸馏版 I2V 增加 480p 的 `/v1/videos` 表单示例；`action generation` 章节改为 "Nano 或 Edge policy 检查点均可单卡服务"，并补充 Edge-Policy-DROID 启动方式。
5. 修订 Parameters 与 `flow_shift` 语义：`flow_shift` 默认值从笼统的 "T2I 3.0 / 视频和 action 10.0" 细化为 "T2I 3.0、非 Edge 视频与所有 action 10.0、Edge 视频 3.0"，并明确蒸馏版忽略请求级 `flow_shift`、`negative-prompt CFG` 与 `num_inference_steps`。

测试、配置与部署配套：无。本 PR 仅改动 1 个 mdx 文档文件，无源码、测试或 CI 配置联动。

关键文件：

- `docs/cookbook/diffusion/Cosmos/Cosmos3.mdx`（模块文档；类别 docs；类型 documentation）：本 PR 唯一变更文件：扩充模型支持表格、新增 Edge 与蒸馏版检查点的启动 / 请求配方，并修订 `flow_shift` 默认值与参数语义，是全部文档价值的载体。

关键符号：未识别

关键源码片段

`docs/cookbook/diffusion/Cosmos/Cosmos3.mdx`

本 PR 唯一变更文件：扩充模型支持表格、新增 Edge 与蒸馏版检查点的启动 / 请求配方，并修订 `flow_shift` 默认值与参数语义，是全部文档价值的载体。

```
# Cosmos3-Edge 是 4B 稠密模型，面向 256p/480p 生成，单卡即可服务
# 其视频默认配置为 832x480 + guidance_scale=5.0，图像默认 640x640 + guidance_scale=7.0
sglang serve \
  --model-path nvidia/Cosmos3-Edge \
  --num-gpus 1

# 蒸馏版 Super 检查点为 64B 模型，建议多卡部署
# SGLang 启动时读取 scheduler/scheduler_config.json 自动识别蒸馏变体
# 并强制使用检查点内置的固定 4 步 sigma 调度与 guidance_scale=1.0
# 因此不要手动调节 num_inference_steps 或 flow_shift
sglang serve \
  --model-path nvidia/Cosmos3-Super-Text2Image-4Step \
  --num-gpus 4
```

评论区精华

该 PR 没有任何 review 评论或讨论线程（`comments_count=0`、`review_comments_count=0`，由作者 mickqian 直接合并），因此没有可见的设计交锋或未解决疑虑。文档中承担关键语义的表述——配置驱动的变体识别、蒸馏版固定 sigma 与强制 `guidance_scale=1.0`——均直接继承自 #31590 的实现决策，未被评审质疑。

- 暂无高价值评论线程

风险与影响

- 风险：本质为零回归风险的纯文档变更。主要风险是文档与实现的漂移：蒸馏版固定 4 步 sigma、强制 guidance_scale=1.0、以及 flow_shift 分模式默认值（Edge 视频 3.0、非 Edge 视频与 action 10.0）都依赖 #31590 引入的启动期配置检测逻辑（如 scheduler/scheduler_config.json 读取），若后续实现调整而文档未同步更新，会直接误导用户部署。此外 --num-gpus 4 仅是建议值，64B 模型在单张超大显存卡上也可运行，文档中的数字不应被理解为硬性要求。
- 影响：对用户：为 Cosmos3 Edge / Edge-Policy-DROID 与两个蒸馏版 Super 检查点提供开箱即用的启动与请求配方（含默认尺寸、guidance、支持尺寸清单），省去试错成本，尤其是蒸馏版不可调调度参数的约束提示。对团队：填补 #31590 功能合入后缺失的用户文档，降低相关支持提问；为后续 diffusion 模型文档提供可复用模板。对系统运行时零影响，不涉及任何代码路径。
- 风险标记：纯文档变更，文档与实现同步风险

关联脉络

- PR #31590 Add Cosmos3 Edge and Distilled checkpoints support: 本 PR 是 #31590 的直接文档后续：后者在代码层新增 Cosmos3 Edge、Edge-Policy-DROID 与 Super 4-step 蒸馏检查点的支持及变体自动识别机制，本 PR 将其固化为 cookbook 部署与请求配方，文档中的关键行为表述均来自该 PR 的实现。