计算加速套件 TACO Kit

TACO DiT 推理加速引擎

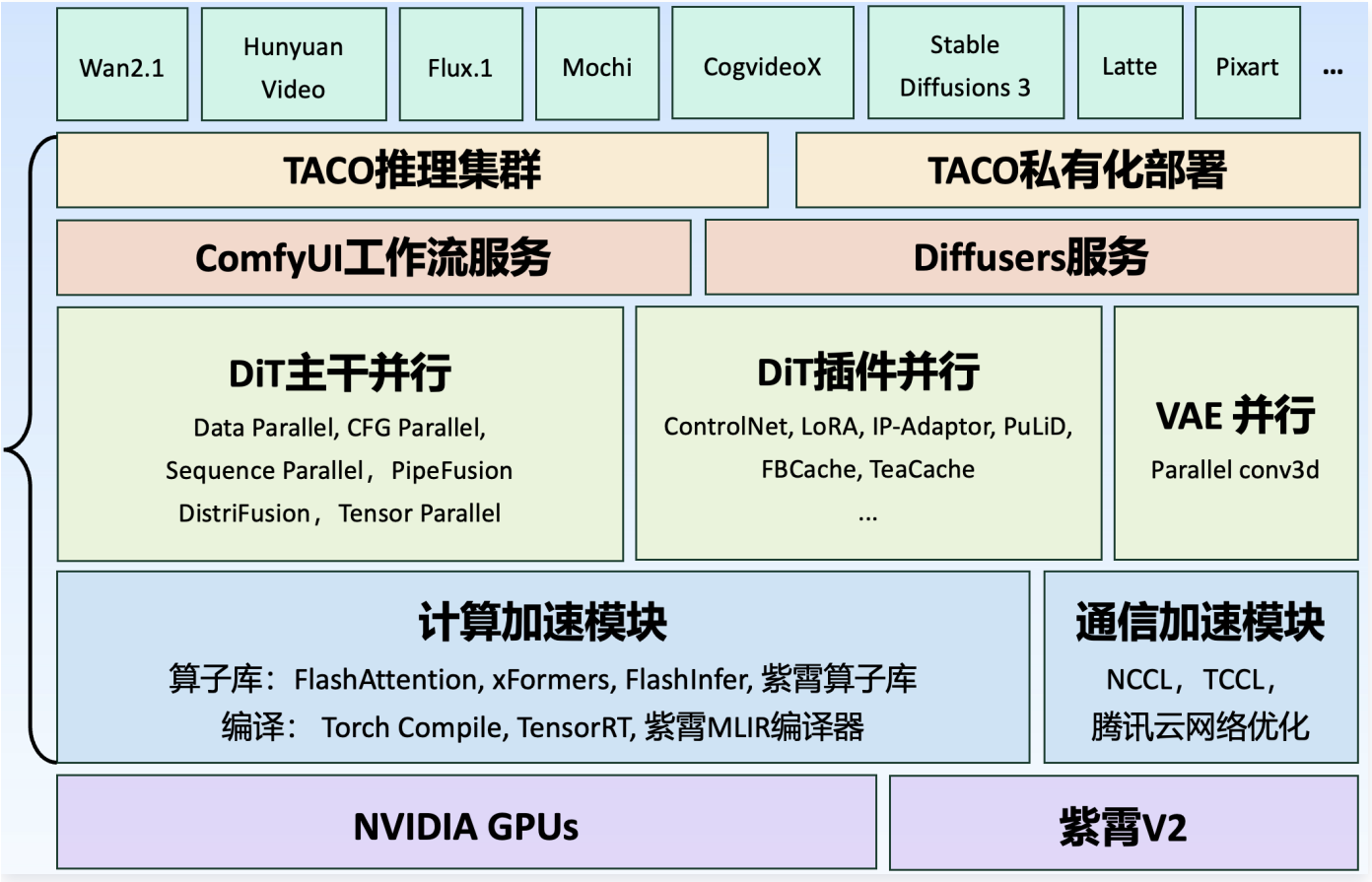
腾讯云

文档目录

TACO DiT 推理加速引擎
TACO DiT 概述
产品简介

最近更新时间：2025-08-12 18:58:02

TACO DiT（TencentCloud Accelerated Computing Optimization DiT）是一款基于腾讯云异构计算产品开发的生图模型（DiT）推理加速引擎，专注于提升文生图、文生视频以及图生视频等场景的推理效率。TACO DiT 支持单卡加速以及业内独特的多卡并行加速能力，在满足客户生成速率的前提下，还能够有效控制生成成本，帮助客户优化业务成本结构。



作为一款可扩展的高性能推理 DiT 引擎，TACO DiT 能够兼容 GPU 以及紫霄硬件，其核心能力概括如下：

- 支持主干网络混合并行

  TACO DiT 提供六种基础并行方法，其中 Data、Sequence、PipeFusion 和 CFG 可以灵活组合使用，以适配底层硬件并扩展至大规模集群。此外，DistriFusion 和 TensorParallel 两种并行策略则作为独立方案运行。
- 支持解码模块 VAE 并行

针对扩散模型后处理中的解码 VAE 模块在生成高分辨率图像时可能出现的 OOM 问题，TACO DiT 开发了 Patch Parallel 版本的 VAE，有效提升了处理效率。

- 支持多种模型

  目前，TACO DiT 已兼容多种主流文生图和视频生成模型，包括 Stable Diffusion、Flux 以及 CogVideoX 等。

- 兼容其他加速模块

  TACO DiT 能够与多种计算加速模块协同工作，例如 FlashAttention、TensorRT 和 torch.compile 等算子及编译加速方案，同时支持在公有云和私有化环境中部署和使用。

应用场景

最近更新时间：2025-08-12 18:58:02

随着当前文生图、文生视频、图生视频技术的快速发展，TACO DiT 的应用场景可分为在线业务和离线业务两类：

TACO DiT 在线业务

- 自媒体内容的生成
  - 短视频创作：为短视频平台创作高质量的视频内容，包括图片素材生成、视频制作、特效制作等。
  - 教育内容制作：为在线教育平台生成教学视频、课件插图和动画，提升学习体验。
  - 品牌宣传：帮助企业生成品牌宣传视频、动态广告图和视觉故事，增强品牌传播效果。
- 影视作品的生成
  - 场景设计与分镜头图：生成电影或电视剧的场景概念图和分镜头设计，优化拍摄规划。
  - 特效视频制作：为影视后期制作提供高质量的特效视频素材，减少人工制作成本。
- 电商商品展示图的制作
  - 虚拟模特与场景生成：生成符合品牌风格的虚拟模特图和商品场景图，提升商品吸引力。
  - 动态商品展示：制作商品的360度旋转视频或动态展示效果，增强用户的购物体验。
  - 促销活动视觉设计：生成节日促销海报、动态广告视频等，降低设计成本，快速响应市场需求。

TACO DiT 离线业务

- 设计图的润色与优化
  - 平面设计优化：为广告、海报、包装设计等提供专业的视觉优化服务，提升设计质量。
  - 建筑设计图优化：对建筑设计图进行细节润色，生成更具吸引力的展示效果。
  - 工业设计支持：优化产品设计图纸，帮助企业更好地展示产品概念。
- 高质量渲染图片的生成
  - 产品设计与展示：生成高精度的产品渲染图，用于产品宣传、用户手册和展示。
  - 建筑可视化：生成建筑项目的 3D 渲染图和虚拟漫游视频，帮助客户更直观地了解设计方案。
  - 汽车与机械设计：生成汽车、机械设备的渲染图，用于设计评审和市场推广。
- 游戏原稿图的生成
  - 游戏角色设计：生成游戏角色的原画、概念图和动态展示，为游戏开发提供创意支持。
  - 场景与地图设计：生成游戏场景、地图概念图和细节设计，提升游戏的视觉表现力。
  - 宣传素材制作：为游戏宣传制作高质量的海报、动态视频和视觉特效，吸引玩家关注。
- 文旅与文化
  - 虚拟旅游体验：生成旅游景点的虚拟漫游视频，吸引游客关注。
  - 文化遗产保护：生成文化遗产的数字化展示图和视频，促进文化传播。

支持模型

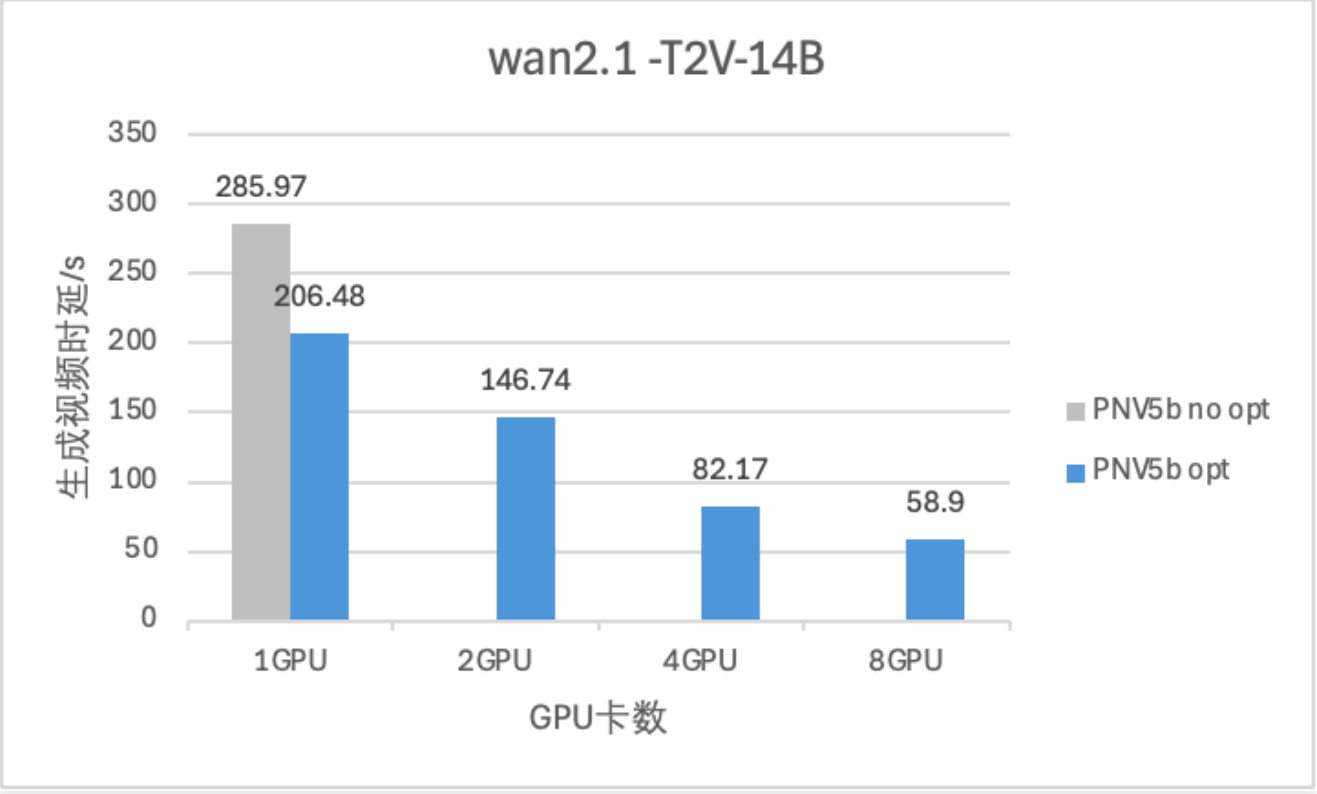
最近更新时间：2025-08-12 18:58:02

目前 TACO DiT 可支持大部分主流模型，满足 AIGC 图像创作需求。TACO DiT 优化节点兼容 ComfyUI 工作流，客户无需改造，可无缝迁移业务。

支持模型

- Flux.1-dev
- Flux.1-kontext-dev
- Wan2.1-T2V-1.3B
- Wan2.1-T2V-14B
- CogVideo-5B

以 Wan2.1-T2V-14B 为例，TACO DiT 优化成果如下：
生成一个640×640×65的视频，20次迭代。TACO DiT 在 PNV5b 的设备上，通过各种优化措施，最终效果如下：



在生成同样效果的视频情况下，TACO DiT 相较于开源版本的视频生成速度，单卡可加速1.38x，双卡加速1.94x，四卡加速3.48x，八卡加速4.85x。

支持特性

最近更新时间：2025-08-12 18:58:02

整体看，DiT 类模型的加速场景主要包括两大类：

- 时延敏感型场景，客户期望单次出图/视频的时间尽可能短，提升用户体验。
- 成本敏感型场景，客户期望尽可能利用有限的算力资源，提升请求服务量。

针对上述两类需求，TACO DiT 分别提供多卡并行和单卡加速两种解决方案，二者可以同时使用。

- 多卡并行加速

  通过聚合多张卡的算力，降低单次推理耗时。主要的并行方法包括：
  - 混合序列并行：Ulysses 并行和 Ring-Attention 并行。文生图/视频场景，通常会将图片进行 patch 处理变成类似文本的序列，再交给 Transformer 进行噪声预测，由于图片的分辨率/信息量较大，导致输入的序列长度较长，Attention 部分计算耗时占比高。为了降低单次推理时延，通过混合序列并行加速计算。
  - CFG（Classifier-Free Guidance）并行：文生图/视频场景，为了提升生成内容的质量和效果，通常会包含有条件（conditioned）的预测和无条件（unconditioned）的预测，从本质上看，二者只是输入序列是否包含文本提示等方面的差异，计算逻辑完全一致，所以可以尝试使用不同的卡同时计算两种预测。
  - PipeFusion 并行：将输入图片/视频切分成多个 patch，通过流水线方式在多个设备之间并行计算，降低推理时延。
- 单卡加速
  - 算子优化：针对计算耗时热点部分，提供定制算子或者量化加速。
  - 计算 Cache 优化：减少重复的计算步骤，降低推理时延。
  - 图编译优化：针对琐碎小算子区域使用编译优化，加速计算图执行。

由于文生图/视频场景工作流的定义较为灵活，所以业务场景的节点多种多样，个性化程度较高，所以推理优化往往需要一些专家介入，并提供优化指导建议，TACO DiT 团队可以提供专业的 技术支持，协助重点客户进行工作流的极致性能优化。

TACO DiT 部署

最近更新时间：2025-08-12 18:58:02

本文主要介绍如何部署和使用 TACO DiT 推理加速服务。

获取 TACO DiT custom 依赖包

外部依赖包和 TACO DiT 工具包都放到 custom node 目录下。

配置外部依赖的开源仓库

```
git clone https://github.com/XLabs-AI/x-flux-comfyui.git
cd x-flux-comfyui
pip install -r requirements.txt

git clone https://github.com/welltop-cn/ComfyUI-TeaCache.git
cd ComfyUI-TeaCache
pip install -r requirements.txt

git clone https://github.com/yolain/ComfyUI-Easy-Use.git
cd ComfyUI-Easy-Use
pip install -r requirements.txt
```

获取并配置 TACO DiT 环境

```
wget ##TACO DiT安装文件##
cd xdit-comfyui
pip install -r requirements.txt
```

TACO DiT 安装文件请联系 技术支持 获取。

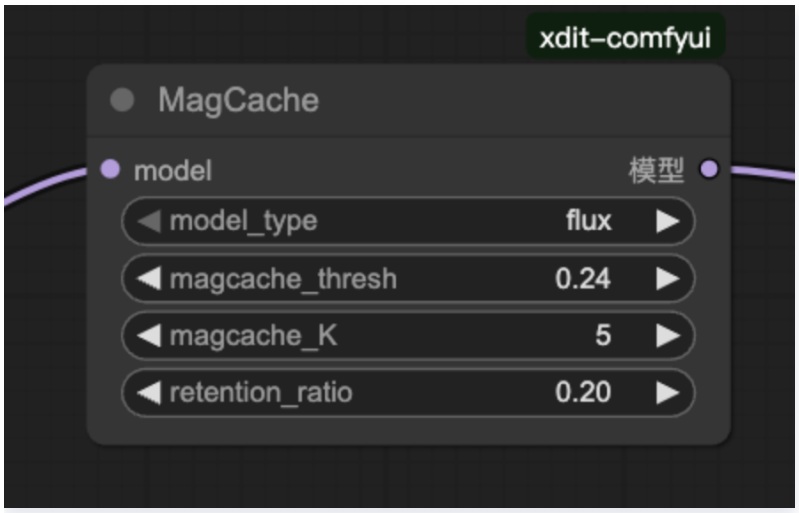
获取并安装 SageAttention

```
git clone https://github.com/thu-ml/SageAttention.git
cd SageAttention
export EXT_PARALLEL=4 NVCC_APPEND_FLAGS="--threads 8" MAX_JOBS=32
# parallel compiling (Optional)
python setup.py install # or pip install -e .
```

客户部署

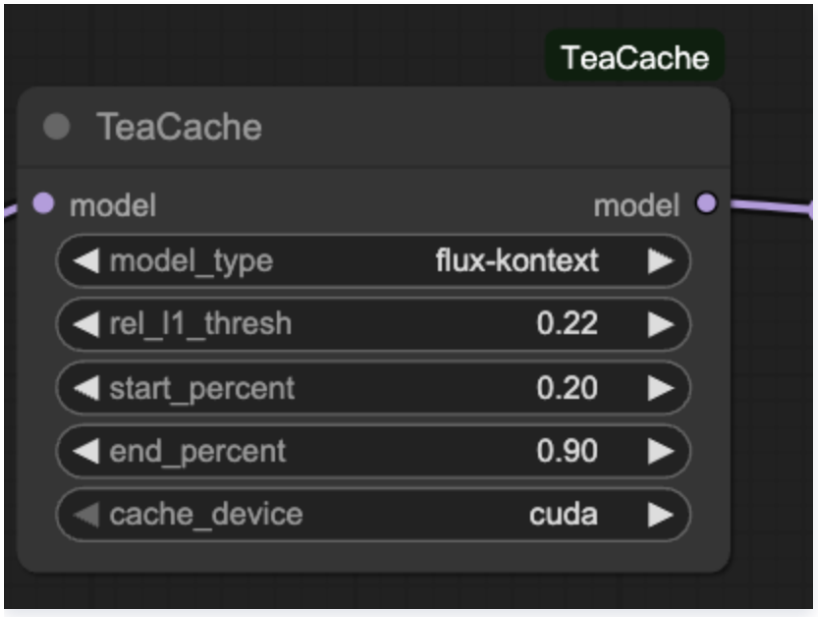
根据不同需求场景，您可以创建不同工作流，主要的 custom node 如下：

MagCache



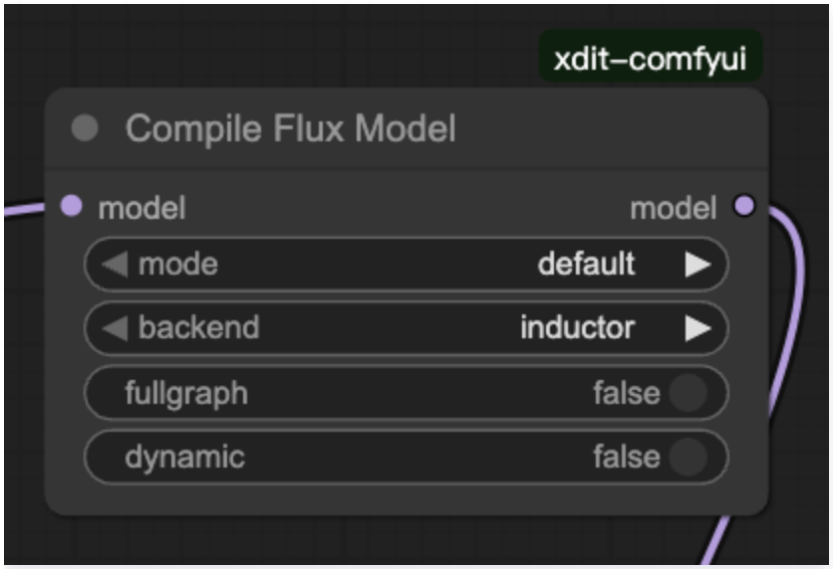
> **说明：**
> 您可修改 model_type 切换不同的模型，如果精度出现下降可以通过降低 magcache_thresh 和 magcache_K 以及增大 retention_ratio 进行优化。

TeaCache



> **说明：**
> 您可修改 model_type 切换不同的模型，如果精度出现下降可以通过降低 rel_l1_thresh 以及调整 start_percent 和 end_percent 进行优化。

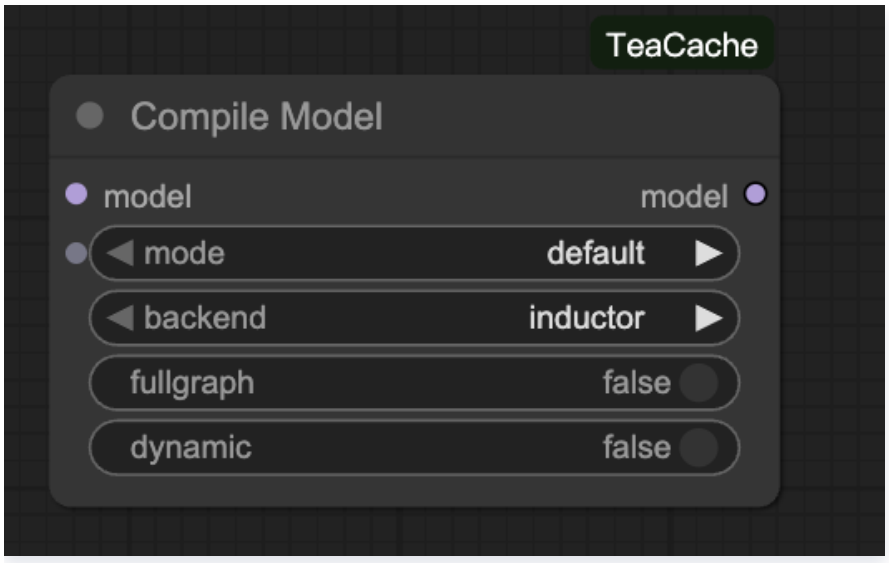
Compile Flux Model



说明：

专为 Flux 模型定制的 Compile Node，第一次执行会对执行流程进行编译，因此会增加第一次的执行时间，但是第二次之后的执行时间会明显下降（**当存在 offload 时可能会编译失败，因此存在 offload 时慎用**）。

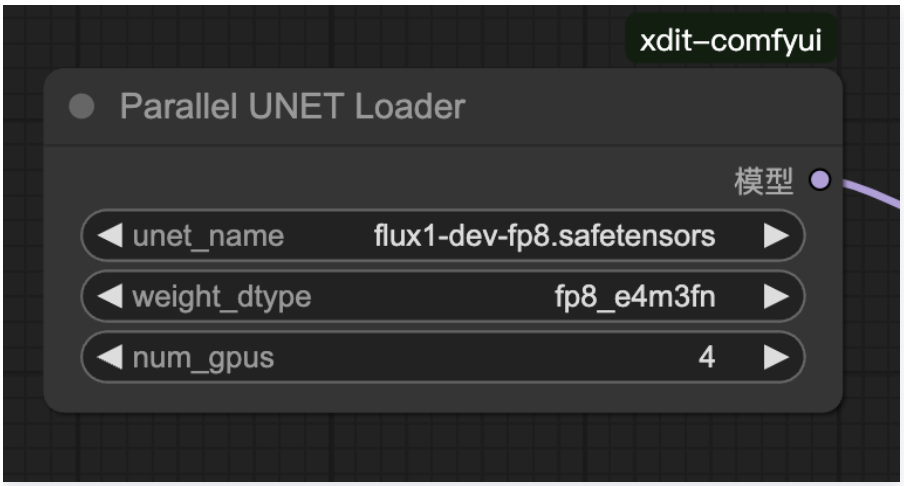
Compile Model



说明：

通用模型的 Compile Node，第一次执行会对执行流程进行编译，因此会增加第一次的执行时间，但是第二次之后的执行时间会明显下降（**当存在 offload 时可能会编译失败，因此存在 offload 时慎用**）。

ParallelUNETLoader

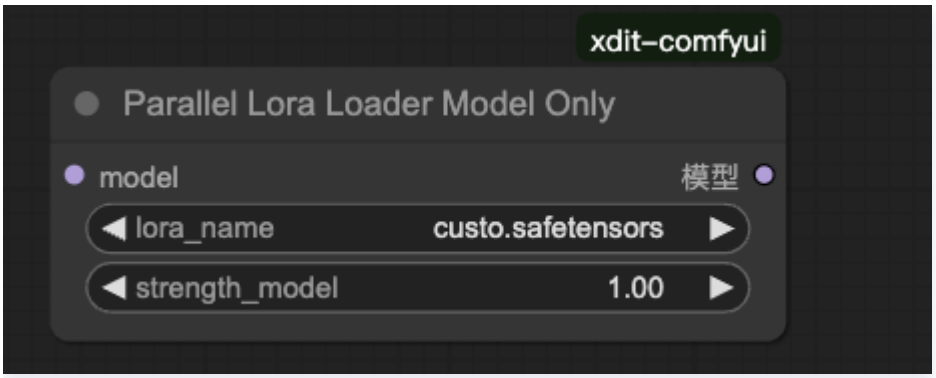


说明：
UNETLoader 的并行版本，通过在启动后设置 num_gpus 可以方便的执行多卡并行（一次 ComfyUI 启动只能支持一种配置，切换卡数需要重启 ComfyUI）。

ParallelLoraLoader

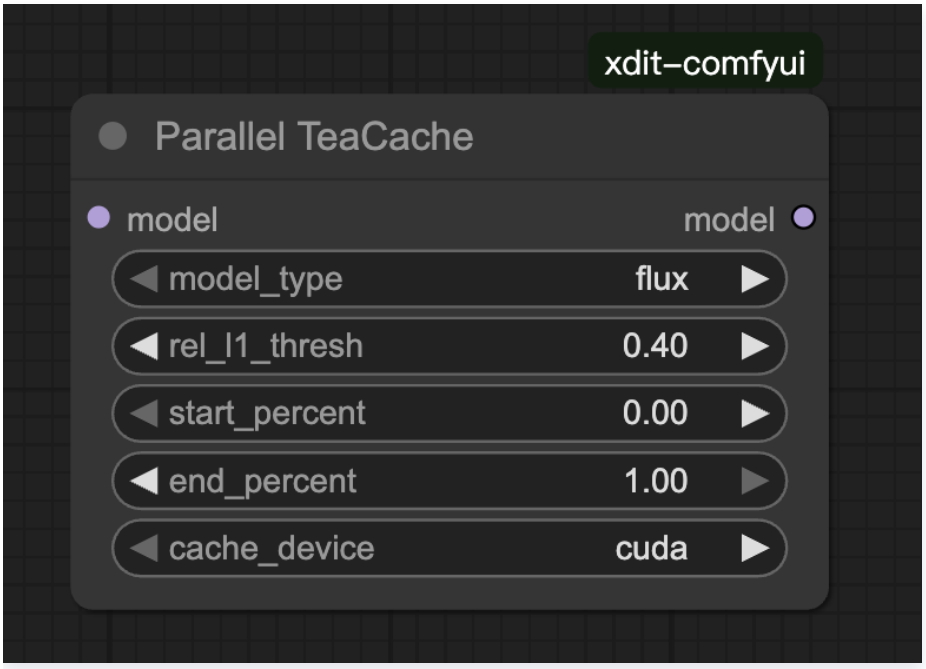
说明：
LoraLoader 的并行版本，与 ParallelUNETLoader 配套使用。

ParallelLoraLoaderModelOnly



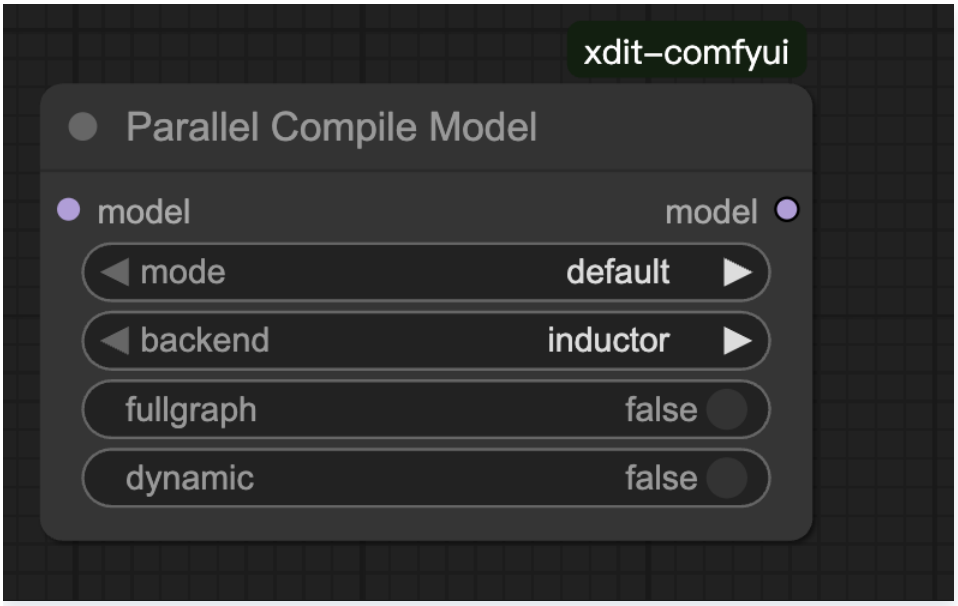
> **说明：**
> LoraLoaderModelOnly 的并行版本，与 ParallelUNETLoader 配套使用。

ParallelTeaCache



> **说明：**
> TeaCache 的并行版本，与 ParallelUNETLoader 配套使用。

ParallelCompileModel



说明：
CompileModel 的并行版本，与 ParallelUNETLoader 配套使用（**当存在 offload 时可能会编译失败，因此存在 offload 时慎用**）。

验证工作流

单卡

单卡执行命令如下：

```
python3 main.py --fast --port 8190 --use-sage-attention
```

为您提供以下样例工作流：

- 单卡Flux.1-dev样例工作流（MagCache）
- 单卡Flux.1-dev样例工作流（TeaCache）
- 单卡Wan2.1-T2V样例工作流

多卡

多卡执行命令如下：

```
python3 main.py --disable-cuda-malloc --fast --port 8190
```

- 多卡Flux.1-dev样例工作流
- 多卡Wan2.1-T2V样例工作流

实践教程
TACO DiT 加速 Flux 模型

最近更新时间：2025-08-12 18:58:02

本文档以 Flux 系列模型为例，演示 TACO DiT 部署和使用流程。

环境准备

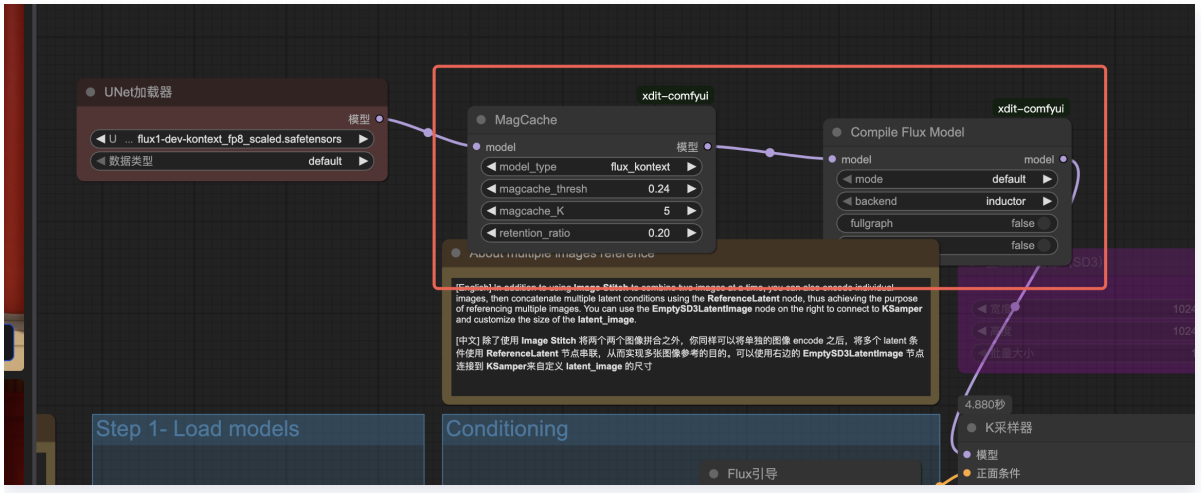
- 机型：GC49d 一卡

> **说明：**
> 如果您的业务场景需要使用当前机型，请联系 技术支持 。

- 软件版本：
  - PyTorch 2.7.1
  - CUDA 12.8

基于 Flux.1-kontext-dev 最佳实践

工作流：Flux.1-kontext-dev
客户按需将工作流放到模型加载器（如果有 lora，放到 lora 加载器之后即可）之后，单击**运行**即可。MegCache 节点和 Compile Flux Model 节点可单独使用也可以配套使用。



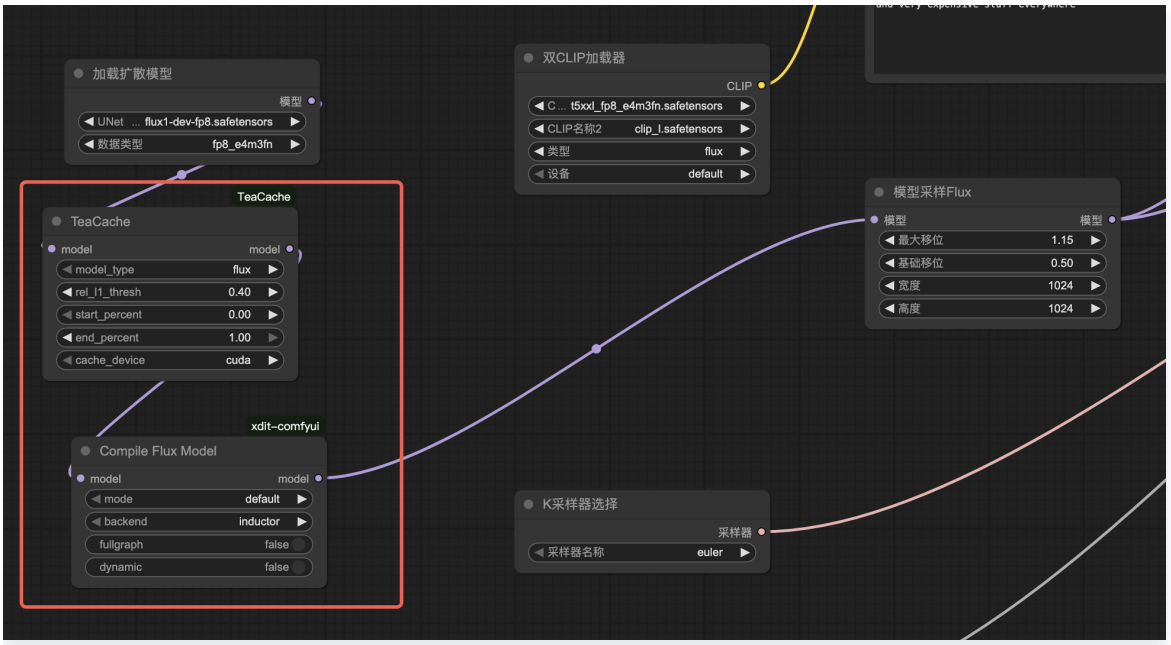
整体生图时延平均提速4.5倍，从23.6s降低到5.2s。运行结果如下：

Baseline	TACO DiT 优化

基于 Flux.1-dev 最佳实践

工作流：Flux.1-dev

客户将上述工作流放置在模型加载器之后，单击**运行**即可。



整体生图时延平均提速3.9倍，从13.9s降低到3.6s。运行结果如下：

Baseline	TACO DiT 优化


13.91s
13.89s


队列
3.55s
3.61s

TACO DiT 加速 Wan2.1-T2V-14B 模型

最近更新时间：2025-08-12 18:58:02

本文档以 Wan2.1-T2V-14B 模型为例，演示 TACO DiT 部署和使用流程。

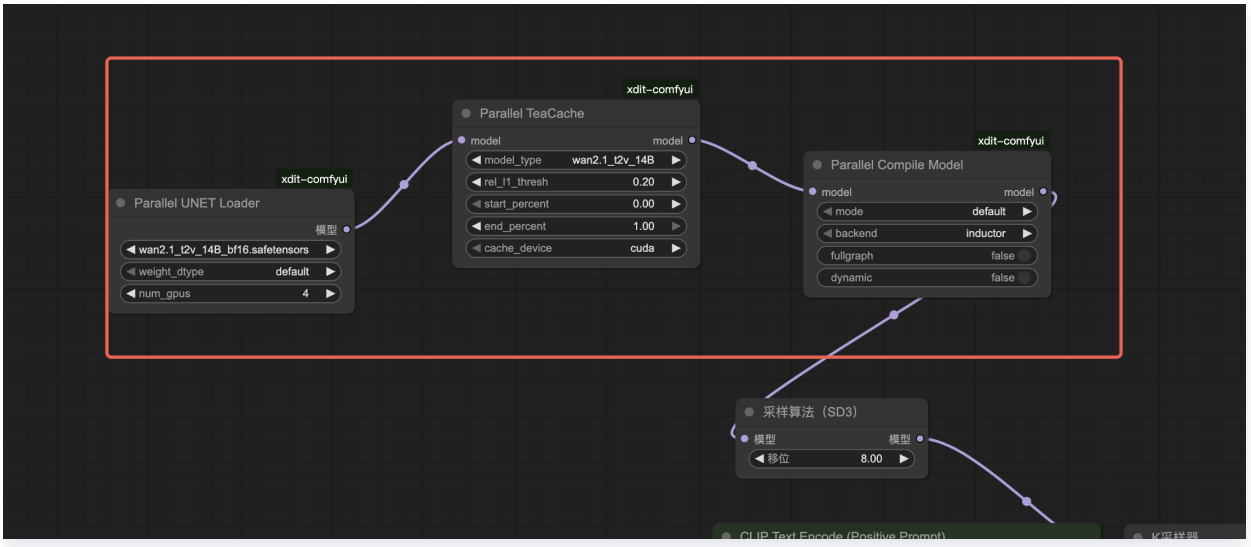
环境准备

- 机型：PNV5b 四卡
- 软件版本：
  - PyTorch 2.7.1
  - CUDA 12.8

基于 Wan2.1-T2V-14B 四卡最佳实践

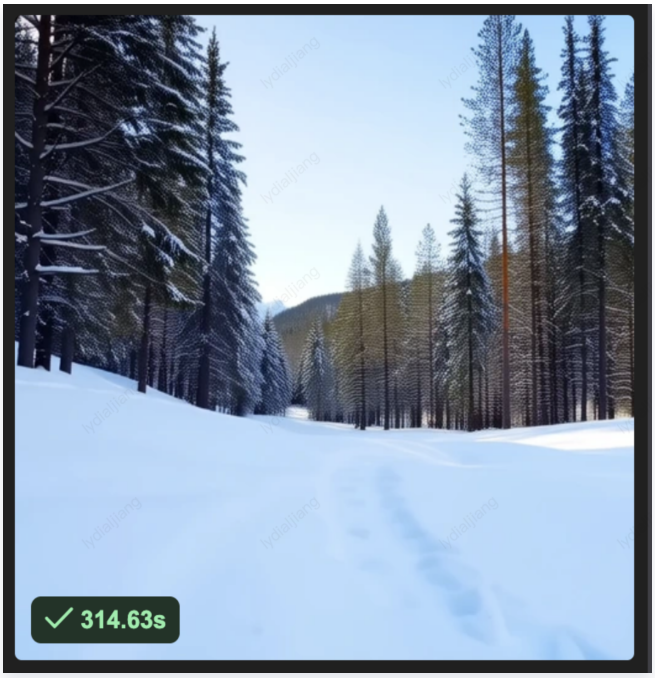
工作流：Wan2.1-T2V-14B

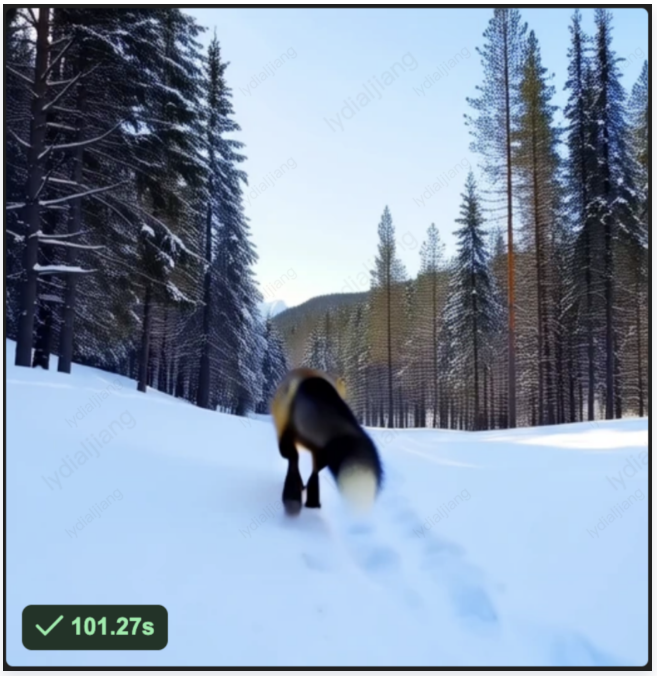
客户将上述工作流放置在模型加载器之后，单击**运行**即可。



整体生图时延平均提速3.1倍，从314.63s降低到101.27s。运行结果如下：

Baseline	TACO DiT 优化


314.63s


101.27s