1.0.0

MindIE 是什么

文档版本	01
发布日期	2025-05-09

华为技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

安全声明

漏洞处理流程

华为公司对产品漏洞管理的规定以“漏洞处理流程”为准，该流程的详细内容请参见如下网址：
https://www.huawei.com/cn/psirt/vul-response-process
如企业客户须获取漏洞信息，请参见如下网址：
https://securitybulletin.huawei.com/enterprise/cn/security-advisory

目 录

1 简介

MindIE（Mind Inference Engine，昇腾推理引擎）是华为昇腾针对AI全场景业务的推理加速套件。通过分层开放AI能力，支撑用户多样化的AI业务需求，使能百模千态，释放昇腾硬件设备算力。向上支持多种主流AI框架，向下对接不同类型昇腾AI处理器，提供多层次编程接口，帮助用户快速构建基于昇腾平台的推理业务。

总体架构

MindIE提供了基于多种AI场景下的推理解决方案，具有强大的性能、健全的生态，帮助用户快速开展业务迁移、业务定制。MindIE架构图如图1-1所示，主要组件介绍如表1-1所示。

图 1-1 昇腾推理引擎架构图

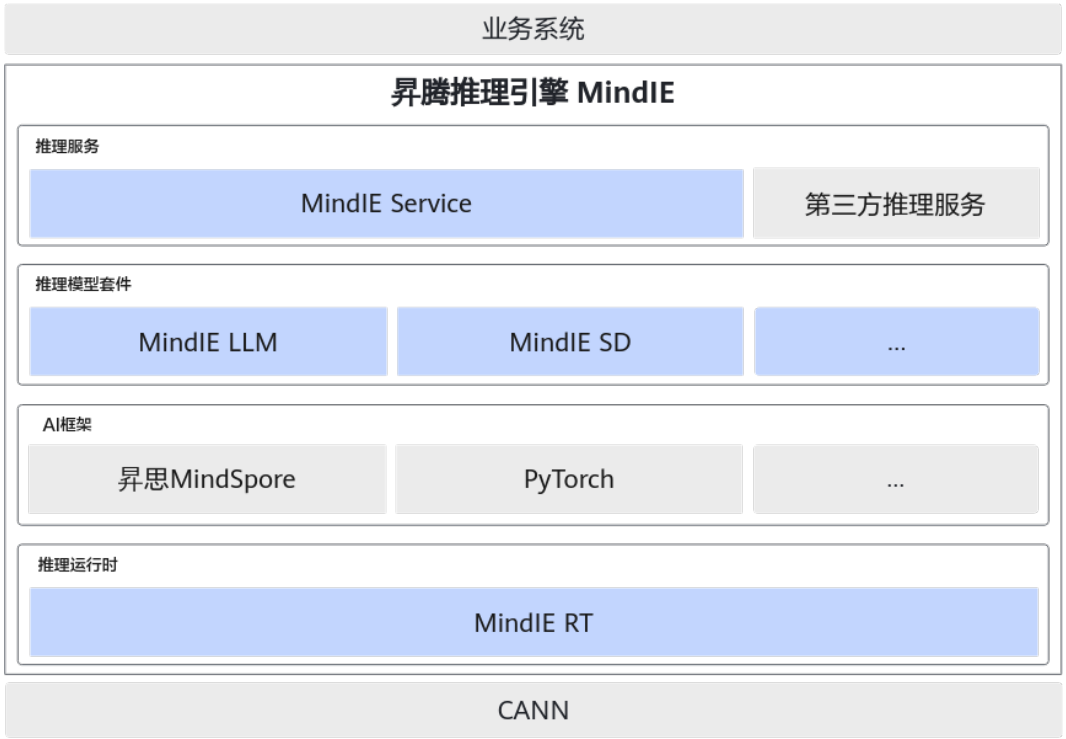


表 1-1 组件介绍

名称	说明
MindIE Service	MindIE Service针对通用模型的推理服务化场景，实现开放、可扩展的推理服务化平台架构，支持对接业界主流推理框架接口，满足大语言模型的高性能推理需求。MindIE Service包含MindIE MS、MindIE Server、MindIE Client和MindIE Benchmark四个子组件。 • MindIE MS：提供服务策略管理和运维能力； • MindIE Server：作为推理服务端，提供模型服务化能力； • MindIE Client：提供服务客户端标准API，简化用户服务调用； • MindIE Benchmark：提供测试大语言模型在不同配置参数下推理性能和精度的能力。 MindIE Service向下调用了MindIE LLM组件能力。
MindIE LLM	MindIE LLM（Mind Inference Engine Large Language Model，大语言模型）是针对大模型优化推理的高性能SDK，包含深度优化的模型库、大模型推理优化器和运行环境，提升大模型推理易用性和性能。
MindIE SD	MindIE SD（Mind Inference Engine Stable Diffusion，Diffusion系列大模型）是MindIE解决方案下的多模态生成推理框架组件，其目标是为多模态生成系列大模型推理任务提供在昇腾硬件及其软件栈上的端到端解决方案，软件系统内部集成各功能模块，对外呈现统一的编程接口。
MindIE Torch	MindIE Torch对接PyTorch框架，提供PyTorch模型推理加速能力。PyTorch框架上训练的模型利用MindIE Torch提供的简易C++/Python接口，少量代码即可完成模型迁移，实现高性能推理。 MindIE Torch向下调用了MindIE RT组件能力。
MindIE RT	MindIE RT（Mind Inference Engine Runtime，推理引擎运行时）基于昇腾异构计算架构CANN提供图引擎能力，当前支持小模型推理场景，更多关于推理运行时能力的构建工作正在进行中，敬请期待。

关键功能特性

- 服务化部署

  提供用户侧接口、调度优化、多模型业务串流等能力。提供模型管理，DevOps等服务化调度能力，详情请参见《**MindIE Service开发指南**》。
- 大模型推理

  提供大模型推理能力，支持大模型业务全流程，逐级能力开放，使能大模型客户需求定制化，详情请参见《**MindIE LLM开发指南**》。
- 多模态生成

  支持SD模型迁移推理，高效实现应用部署，场景化落地SD应用，满足客户精度及性能要求，详情请参见《**MindIE SD开发指南**》。
- PyTorch模型迁移（当前能力构建中）

  对接主流PyTorch框架，实现训练到推理的平滑迁移，提供通用的图优化并行推理能力，提供用户深度定制优化能力，详情请参见《**MindIE Torch开发指南**》。

快速安装 MindIE

安装MindIE请参考《**MindIE安装指南**》。

2 支持的模型列表

2.1 大语言模型列表

须知

以下模型需配合ATB Models模型库使用，ATB Models的安装方式请参见《MindIE安装指南》中“安装开发环境 > **安装ATB Models**”章节。

MindIE支持的大语言模型列表如下所示。

Baichuan 系列

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
Baichuan2-7B	• Atlas 800I A2 推理产品：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1。	• FP16：支持 • BF16：不支持	• W8A8量化：仅Atlas 800I A2 推理产品支持 • W8A16量化：仅Atlas 800I A2 推理产品支持	• MindIE Service：支持 • TGI：支持	不支持	**链接**

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
Baichuan2-13B	• Atlas 800I A2 推理产品：支持的卡数为2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1或2。	• FP16：支持 • BF16：不支持	• W8A8量化：仅Atlas 800I A2 推理产品支持 • W8A16量化：仅Atlas 800I A2 推理产品支持	• MindIE Service：支持 • TGI：支持	不支持	链接

Bloom 系列

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
Bloom-7B	• Atlas 800I A2 推理产品：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为4。	• FP16：支持 • BF16：不支持	不支持	不支持	不支持	链接
Bloom-176B	• Atlas 800I A2 推理产品：支持的卡数为8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：支持 • BF16：不支持	W8A16量化：仅Atlas 800I A2 推理产品支持	不支持	不支持	链接

ChatGLM 系列

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
ChatGLM2-6B（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1、2或4。	• FP16：支持 • BF16：不支持	• W8A8量化：仅Atlas 800I A2 推理产品支持 • 稀疏量化：仅Atlas 300I Duo 推理卡支持	• MindIE Service：支持 • TGI：支持	不支持	链接
ChatGLM3-6B	• Atlas 800I A2 推理产品（32G）：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1或2。	• FP16：支持 • BF16：不支持	不支持	• MindIE Service：支持 • TGI：不支持	不支持	链接
ChatGLM3-6B-32K	• Atlas 800I A2 推理产品：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1或2。	• FP16：支持 • BF16：不支持	W8A8量化：仅Atlas 800I A2 推理产品支持	• MindIE Service：支持 • TGI：不支持	不支持	链接

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
GLM4-9B-Chat	• Atlas 800I A2 推理产品：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1或2。	• FP16：支持 • BF16：仅Atlas 800I A2 推理产品支持	• W8A8量化：仅Atlas 800I A2 推理产品支持 • 稀疏量化：仅Atlas 300I Duo 推理卡支持 • W8A8C8量化：仅Atlas 800I A2 推理产品（32G）支持	• MindIE Service：支持 • TGI：不支持	Atlas 800I A2 推理产品（64G）支持的长度最长为1M	链接

CodeLLaMA 系列

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
CodeLLaMA-13B	• Atlas 800I A2 推理产品（32G）：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：1、2或4	• FP16：支持 • BF16：仅Atlas 800I A2 推理产品支持	不支持	• MindIE Service：支持 • TGI：不支持	不支持	链接

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
CodeLLaMA-34B	• Atlas 800I A2 推理产品（32G）：支持的卡数为4或8。 • Atlas 300I Duo 推理卡：支持的卡数为2或4。	• FP16：支持 • BF16：仅Atlas 800I A2 推理产品支持	• W8A8量化：仅Atlas 800I A2 推理产品支持 • 稀疏量化：仅Atlas 300I Duo 推理卡支持	• MindIE Service：支持 • TGI：不支持	不支持	链接
CodeLLaMA-70B	• Atlas 800I A2 推理产品（32G）：支持的卡数为8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：支持 • BF16：仅Atlas 800I A2 推理产品支持	不支持	不支持	不支持	链接

DeepSeek 系列

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
DeepSeek-Coder-6.7B	• Atlas 800I A2 推理产品（32G）：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1或2。	• FP16：支持 • BF16：仅Atlas 800I A2 推理产品支持	不支持	不支持	不支持	链接

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
DeepSeek-Coder-7B	• Atlas 800I A2 推理产品（32G）：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1或2。	• FP16：支持 • BF16：仅Atlas 800I A2 推理产品支持	不支持	不支持	不支持	链接
DeepSeek-Coder-33B	• Atlas 800I A2 推理产品（32G）：支持的卡数为4或8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：仅Atlas 800I A2 推理产品支持	不支持	不支持	不支持	链接
DeepSeek-MoE-16B	• Atlas 800I A2 推理产品（32G）：支持的卡数为4或8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：不支持	不支持	• MindIE Service：支持 • TGI：不支持	不支持	链接
DeepSeek-V2-Lite-16B	• Atlas 800I A2 推理产品：支持的卡数为2、4或8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：不支持	W8A16量化：仅Atlas 800I A2 推理产品支持	不支持	不支持	链接

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
DeepSeek-V2-236B	• Atlas 800I A2 推理产品（64G）：支持的卡数为16。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：不支持	W8A16量化：仅Atlas 800I A2 推理产品支持	不支持	不支持	链接

InternLM 系列

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
InternLM-20B （该模型预计下个版本将日落）	• Atlas 800I A2 推理产品：支持的卡数为2、4或8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：不支持	不支持	不支持	不支持	链接
InternLM2-20B	• Atlas 800I A2 推理产品：支持的卡数为2、4或8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：不支持	不支持	• MindIE Service：支持 • TGI：不支持	Atlas 800I A2 推理产品（64G）支持的长度最长为200K	链接

LLaMA 系列

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
LLaMA-7B（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1、2或4。	• FP16：支持 • BF16：不支持	不支持	不支持	不支持	链接
LLaMA-13B（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1、2或4。	• FP16：支持 • BF16：不支持	不支持	不支持	不支持	链接
LLaMA-33B（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品：支持的卡数为4或8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：仅Atlas 800I A2 推理产品支持	不支持	不支持	不支持	链接

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
LLaMA-65B（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品（32G）：支持的卡数为8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：仅Atlas 800I A2 推理产品支持	W8A16量化：仅Atlas 800I A2 推理产品支持	• MindIE Service：支持 • TGI：支持	不支持	链接
LLaMA2-7B（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1、2或4。	• FP16：支持 • BF16：不支持	• W8A8量化：仅Atlas 800I A2 推理产品支持 • 稀疏量化：仅Atlas 300I Duo 推理卡支持	• MindIE Service：支持 • TGI：支持	不支持	链接

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
LLaMA2-13B（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1、2或4。	• FP16：支持 • BF16：不支持	• W8A8量化：仅Atlas 800I A2 推理产品支持 • 稀疏量化：仅Atlas 300I Duo 推理卡支持	• MindIE Service：支持 • TGI：支持	不支持	链接
LLaMA2-70B（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品：支持的卡数为8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：仅Atlas 800I A2 推理产品支持	• W8A8量化：仅Atlas 800I A2 推理产品支持 • W8A16量化：仅Atlas 800I A2 推理产品支持	• MindIE Service：不支持 • TGI：支持	不支持	链接

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
LLaMA3-8B	• Atlas 800I A2 推理产品（32G）：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1、2或4。	• FP16：支持 • BF16：仅Atlas 800I A2 推理产品支持	不支持	• MindIE Service：支持 • TGI：不支持	不支持	链接
LLaMA3-70B	• Atlas 800I A2 推理产品：支持的卡数为8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：仅Atlas 800I A2 推理产品支持	W8A16量化：仅Atlas 800I A2 推理产品支持	• MindIE Service：支持 • TGI：不支持	不支持	链接
LLaMA3.1-8B	• Atlas 800I A2 推理产品：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1、2或4。	• FP16：支持 • BF16：仅Atlas 800I A2 推理产品支持	W8A8量化：仅Atlas 800I A2 推理产品支持	• MindIE Service：支持 • TGI：不支持	不支持	链接

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
LLaMA3.1-70B	• Atlas 800I A2 推理产品：支持的卡数为8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：仅Atlas 800I A2 推理产品支持	• kv cache量化：仅Atlas 800I A2 推理产品支持 • Attention量化：仅Atlas 800I A2 推理产品支持	• MindIE Service：支持 • TGI：不支持	Atlas 800I A2 推理产品（64G）支持的长度最长为128K	链接
LLaMA3.1-405B	• Atlas 800I A2 推理产品（64G）：支持的卡数为16或32。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：仅Atlas 800I A2 推理产品支持	不支持	不支持	不支持	链接

Mixtral 系列

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
Mixtral-8x7B-Instruct-V0.1	• Atlas 800I A2 推理产品：支持的卡数为8。 • Atlas 300I Duo 推理卡：支持的卡数为4。	• FP16：支持 • BF16：不支持	W8A8量化：仅Atlas 800I A2推理产品支持	不支持	不支持	链接
Mixtral-8x22B-Instruct-V0.1	• Atlas 800I A2 推理产品（64G）：支持的卡数为8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2推理产品支持 • BF16：不支持	不支持	不支持	不支持	链接

OpenBMB/MiniCPM 系列

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
OpenBMB/MiniCPM-1B-sft-bf16	• Atlas 800I A2 推理产品：不支持。 • Atlas 300I Duo 推理卡：支持的卡数为1。	• FP16：支持 • BF16：不支持	不支持	不支持	不支持	链接
OpenBMB/MiniCPM-2B-sft-bf16	• Atlas 800I A2 推理产品：不支持。 • Atlas 300I Duo 推理卡：支持的卡数为1。	• FP16：支持 • BF16：不支持	不支持	不支持	不支持	链接

Qwen 系列

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
Qwen-7B（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1。	• FP16：支持 • BF16：不支持	W8A8量化：仅Atlas 800I A2推理产品支持	• MindIE Service：支持 • TGI：不支持	不支持	链接
Qwen-14B（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品：支持的卡数为2、4或8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：不支持	W8A8量化：仅Atlas 800I A2推理产品支持	• MindIE Service：支持 • TGI：不支持	不支持	链接
Qwen-72B（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品（32G）：支持的卡数为8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：不支持	W8A16量化：仅Atlas 800I A2推理产品支持	• MindIE Service：支持 • TGI：不支持	不支持	链接

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
Qwen1.5-7B	• Atlas 800I A2 推理产品（32G）：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1。	• FP16：支持 • BF16：不支持	不支持	• MindIE Service：支持 • TGI：不支持	不支持	链接
Qwen1.5-14B	• Atlas 800I A2 推理产品（32G）：支持的卡数为2、4或8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：不支持	W8A8量化：仅Atlas 800I A2 推理产品支持	• MindIE Service：支持 • TGI：不支持	不支持	链接
Qwen1.5-32B	• Atlas 800I A2 推理产品（32G）：支持的卡数为4或8。 • Atlas 300I Duo 推理卡：不支持	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：不支持	W8A8量化：仅Atlas 800I A2 推理产品支持	• MindIE Service：支持 • TGI：不支持	不支持	链接

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
Qwen1.5-72B	• Atlas 800I A2 推理产品：支持的卡数为8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：不支持	W8A16量化：仅Atlas 800I A2 推理产品支持	• MindIE Service：支持 • TGI：不支持	不支持	链接
Qwen1.5-110B	• Atlas 800I A2 推理产品：支持的卡数为8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：不支持	W8A16量化：仅Atlas 800I A2 推理产品支持	不支持	不支持	链接
Qwen2-57B-A14B	• Atlas 800I A2 推理产品：支持的卡数为8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：不支持 • BF16：仅Atlas 800I A2 推理产品支持	不支持	• MindIE Service：支持 • TGI：不支持	不支持	链接

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
Qwen2-7B	• Atlas 800I A2 推理产品：支持的卡数为1、2或4。 • Atlas 300I Duo 推理卡：支持的卡数为1。	• FP16：仅Atlas 300I Duo 推理卡支持 • BF16：仅Atlas 800I A2 推理产品支持	W8A8量化：仅Atlas 800I A2 推理产品支持	• MindIE Service：支持 • TGI：不支持	不支持	链接

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
Qwen2-72B	• Atlas 800I A2 推理产品：支持的卡数为4或8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：不支持 • BF16：仅Atlas 800I A2 推理产品支持	• W8A8量化：仅Atlas 800I A2 推理产品支持 • W8A16量化：仅Atlas 800I A2 推理产品支持 • KV cache量化：仅Atlas 800I A2 推理产品支持	• MindIE Service：支持 • TGI：不支持	Atlas 800I A2 推理产品（64G）支持的长度最长为128K	链接

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
Qwen2.5-7B	• Atlas 800I A2 推理产品：支持的卡数为1或2。 • Atlas 300I Duo 推理卡：支持的卡数为1。	• FP16：仅Atlas 300I Duo 推理卡支持 • BF16：仅Atlas 800I A2 推理产品支持	不支持	• MindIE Service：支持 • TGI：不支持	不支持	链接
Qwen2.5-14B	• Atlas 800I A2 推理产品：支持的卡数为2。 • Atlas 300I Duo 推理卡：1。	• FP16：仅Atlas 300I Duo 推理卡支持 • BF16：仅Atlas 800I A2 推理产品支持	不支持	• MindIE Service：支持 • TGI：不支持	不支持	链接
Qwen2.5-32B	• Atlas 800I A2 推理产品：支持的卡数为4或8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：不支持 • BF16：仅Atlas 800I A2 推理产品支持	W8A8量化：仅Atlas 800I A2 推理产品支持	• MindIE Service：支持 • TGI：不支持	不支持	链接

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
Qwen2.5-72B	• Atlas 800I A2 推理产品：支持的卡数为8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：不支持 • BF16：仅Atlas 800I A2 推理产品支持	W8A8量化：仅Atlas 800I A2 推理产品支持	• MindIE Service：支持 • TGI：不支持	不支持	链接

StarCoder 系列

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
StarCoder-15.5B（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品：支持的卡数为8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：不支持	W8A8量化：支持	• MindIE Service：支持 • TGI：支持	不支持	链接

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
StarCoder2-15B（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品（32G）：支持的卡数为4。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：不支持	W8A8量化：仅Atlas 800I A2 推理产品支持	• MindIE Service：支持 • TGI：不支持	不支持	链接

Vicuna 系列

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
Vicuna-7B（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品（32G）：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1、2或4。	• FP16：支持 • BF16：不支持	不支持	不支持	不支持	链接
Vicuna-13B（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品（32G）：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1、2或4。	• FP16：支持 • BF16：不支持	不支持	不支持	不支持	链接

Yi 系列

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
Yi-6B-200K	• Atlas 800I A2 推理产品（32G）：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1、2或4。	• FP16：支持 • BF16：仅Atlas 800I A2 推理产品支持	不支持	不支持	不支持	链接
Yi-34B-200K（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品（32G）：支持的卡数为4或8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：仅Atlas 800I A2 推理产品支持	不支持	不支持	Atlas 800I A2 推理产品（64G）支持的长度最长为200K	链接

其他系列

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
Bloomz-7B1-MT	• Atlas 800I A2 推理产品（32G）：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：支持 • BF16：不支持	不支持	不支持	不支持	链接

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
CodeGeeX2-6B（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1或2。	• FP16：支持 • BF16：不支持	W8A8量化：仅Atlas 800I A2 推理产品支持	• MindIE Service：支持 • TGI：支持	不支持	链接
CodeShell-7B（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品（32G）：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1。	• FP16：支持 • BF16：不支持	不支持	不支持	不支持	链接
Command-R-Plus-104B	• Atlas 800I A2 推理产品：支持的卡数为8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：支持 • BF16：不支持	不支持	不支持	不支持	链接
Gemma-7B（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品（32G）：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1、2或4。	• FP16：支持 • BF16：仅Atlas 800I A2 推理产品支持	W8A8量化：仅Atlas 800I A2 推理产品支持	不支持	不支持	链接
GPT-NEOX-20B	• Atlas 800I A2 推理产品：支持的卡数为2、4或8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：不支持	不支持	不支持	不支持	链接

模型名称	多卡能力	数据类型	量化	服务化	长序列	模型权重链接
Mistral-7B	• Atlas 800I A2 推理产品（32G）：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1或2。	• FP16：支持 • BF16：不支持	不支持	不支持	不支持	链接
TeleChat12B-V2	• Atlas 800I A2 推理产品：不支持。 • Atlas 300I Duo 推理卡：支持的卡数为2或4。	• FP16：支持 • BF16：不支持	稀疏量化：仅Atlas 300I Duo 推理卡支持	不支持	不支持	链接
Ziya-Coding-34B（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品（32G）：支持的卡数为4或8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2 推理产品支持 • BF16：不支持	不支持	不支持	不支持	链接

2.2 嵌入模型支持列表

说明

- 嵌入模型的具体使用方法请参见《MindIE开源第三方服务化框架适配开发指南》的“TEI > 使用差异 > 模型准备 > **选择MindIE Torch作为模型后端**”章节。
- 由于MindIE服务化框架将于下个版本（版本号：2.0.RC1）起停止对开源第三方服务化框架的默认对接支持，所以嵌入模型将于下个版本（版本号：2.0.RC1）同步下架。

模型名称	多卡能力	数据类型	服务化	长序列	模型权重链接
bge-large-zh-v1.5	• Atlas 800I A2 推理产品（32G）：支持的卡数为1。 • Atlas 300I Duo 推理卡：支持的卡数为1。	• FP16：支持 • BF16：不支持	不支持	不支持	链接

模型名称	多卡能力	数据类型	服务化	长序列	模型权重链接
bge-reranker-large	• Atlas 800I A2 推理产品（32G）：支持的卡数为1。 • Atlas 300I Duo 推理卡：支持的卡数为1。	• FP16：支持 • BF16：不支持	不支持	不支持	链接
bge-m3	• Atlas 800I A2 推理产品（32G）：支持的卡数为1。 • Atlas 300I Duo 推理卡：支持的卡数为1。	• FP16：支持 • BF16：不支持	不支持	不支持	链接

2.3 多模态理解模型列表

须知

- 以下模型需配合ATB Models模型库使用，ATB Models的安装方式请参见《MindIE安装指南》中“安装开发环境 > **安装ATB Models**”章节；ATB Models的使用场景请参见《MindIE LLM开发指南》的“模型推理使用流程 > **ATB Models使用**”章节。
- 当前版本多模态理解模型不支持量化。

模型名称	多卡能力	数据类型	服务化	模型权重链接
LLava-1.6-mistral-7B （该模型预计下个版本将日落）	• Atlas 800I A2 推理产品（64G）：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1、2或4。	• FP16：支持 • BF16：仅Atlas 800I A2推理产品支持	MindIE Service：支持	链接
LLava-1.6-vicuna-7B （该模型预计下个版本将日落）	• Atlas 800I A2 推理产品（64G）：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1、2或4。	• FP16：支持 • BF16：仅Atlas 800I A2推理产品支持	MindIE Service：支持	链接

模型名称	多卡能力	数据类型	服务化	模型权重链接
LLava-1.6-vicuna-13B（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品（64G）：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1、2或4。	• FP16：支持 • BF16：仅Atlas 800I A2推理产品支持	MindIE Service：支持	链接
LLava-v1.6-34b-hf（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品（64G）：支持的卡数为4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1、2或4。	• FP16：支持 • BF16：仅Atlas 800I A2推理产品支持	MindIE Service：支持	链接
LLava-next-video-34b（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品（64G）：支持的卡数为4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1、2或4。	• FP16：支持 • BF16：仅Atlas 800I A2推理产品支持	MindIE Service：支持	链接
LLava-next-video-7b（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品（64G）：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1、2或4。	• FP16：支持 • BF16：仅Atlas 800I A2推理产品支持	MindIE Service：支持	链接
LLava-v1.5-13B（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品（64G）：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1、2或4。	• FP16：支持 • BF16：仅Atlas 800I A2推理产品支持	MindIE Service：支持	链接
LLava-v1.5-7B（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品（64G）：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1、2或4。	• FP16：支持 • BF16：仅Atlas 800I A2推理产品支持	MindIE Service：支持	链接
Qwen-VL-9.6B	• Atlas 800I A2 推理产品（32G）：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：支持的卡数为1、2或4。	• FP16：支持 • BF16：仅Atlas 800I A2推理产品（32G）支持	MindIE Service：支持	链接

模型名称	多卡能力	数据类型	服务化	模型权重链接
Qwen2-Audio-7B-Instruct	• Atlas 800I A2 推理产品：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2推理产品支持 • BF16：仅Atlas 800I A2推理产品支持	MindIE Service：支持	链接
Qwen2-VL-7B	• Atlas 800I A2 推理产品：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2推理产品支持 • BF16：仅Atlas 800I A2推理产品支持	MindIE Service：支持	链接
internVL-Chat-V1-2（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品：支持的卡数为4或8。 • Atlas 300I Duo 推理卡：支持的卡数为2或4。	• FP16：支持 • BF16：仅Atlas 800I A2推理产品支持	MindIE Service：支持	链接
internVL-Chat-V1-5（该模型预计下个版本将日落）	• Atlas 800I A2 推理产品（64G）：支持的卡数为4、8。 • Atlas 300I Duo 推理卡：支持的卡数为2、4。	• FP16：支持 • BF16：不支持	MindIE Service：支持	链接
InternVL2-8B	• Atlas 800I A2 推理产品：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2推理产品支持 • BF16：不支持	MindIE Service：支持	链接
InternVL2-40B	• Atlas 800I A2 推理产品：支持的卡数为4、8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：仅Atlas 800I A2推理产品支持 • BF16：仅Atlas 800I A2推理产品（64G）支持	MindIE Service：支持	链接

模型名称	多卡能力	数据类型	服务化	模型权重链接
MiniCPM-Llama3-V-2_5 （该模型预计下个版本将日落）	• Atlas 800I A2 推理产品（64G）：不支持。 • Atlas 300I Duo 推理卡：支持的卡数为1。	• FP16：支持 • BF16：不支持	不支持	链接
MiniCPM-V-2 （该模型预计下个版本将日落）	• Atlas 800I A2 推理产品（64G）：不支持。 • Atlas 300I Duo 推理卡：支持的卡数为1。	• FP16：支持 • BF16：不支持	不支持	链接
internLM-xcomposer2-vl-7B （该模型预计下个版本将日落）	• Atlas 800I A2 推理产品（64G）：支持的卡数为1。 • Atlas 300I Duo 推理卡：支持的卡数为1。	• FP16：支持 • BF16：不支持	不支持	链接
internLM-XComposer2-4KHD-7B （该模型预计下个版本将日落）	• Atlas 800I A2 推理产品（64G）：支持的卡数为1。 • Atlas 300I Duo 推理卡：支持的卡数为1。	• FP16：支持 • BF16：不支持	不支持	链接

2.4 多模态生成模型列表

说明

模型推理详情请参见《MindIE SD开发指南》。

表 2-1 模型列表

模型名称	多卡能力	数据类型	量化	prompts（提示词）最大长度	模型权重链接
CogView3	• Atlas 800I A2 推理产品（64G）：支持的卡数为1。 • Atlas 300I Duo 推理卡：不支持。	• FP16：不支持 • BF16：支持	不支持	224	链接
DiT	• Atlas 800I A2 推理产品（32G）：支持的卡数为1。 • Atlas 300I Duo 推理卡：支持的卡数为1。	• FP16：支持 • BF16：不支持	不支持	不涉及	链接 链接 链接 链接
HunyuanDiT	• Atlas 800I A2 推理产品：支持的卡数为1。 • Atlas 300I Duo 推理卡：不支持。	• FP16：支持 • BF16：不支持	不支持	256	链接
OpenSora v1.2	• Atlas 800I A2 推理产品（32G）：支持的卡数为1、2、4或8。 • Atlas 300I Duo 推理卡：不支持。	• FP16：支持 • BF16：支持	不支持	300	链接 链接 链接 链接
sd-webui	• Atlas 800I A2 推理产品（32G）：支持的卡数为1。 • Atlas 300I Duo 推理卡：支持的卡数为1。	• FP16：支持 • BF16：不支持	不涉及	不涉及	无
Stable Diffusion 1.5	• Atlas 800I A2 推理产品（32G）：支持的卡数为1或2。 • Atlas 300I Duo 推理卡：支持的卡数为1，可双芯并行。	• FP16：支持 • BF16：不支持	不支持	77	链接
Stable Diffusion 2.1	• Atlas 800I A2 推理产品（32G）：支持的卡数为1或2。 • Atlas 300I Duo 推理卡：支持的卡数为1，可双芯并行。	• FP16：支持 • BF16：不支持	不支持	77	链接

模型名称	多卡能力	数据类型	量化	prompts（提示词）最大长度	模型权重链接
Stable Diffusion XL	• Atlas 800I A2 推理产品（32G）：支持的卡数为1或2。 • Atlas 300I Duo 推理卡：支持的卡数为1，可双芯并行。	• FP16：支持 • BF16：不支持	W8A8量化：Atlas 800I A2 推理产品支持	77	链接
Stable Diffusion XL_controlnet	• Atlas 800I A2 推理产品（32G）：支持的卡数为1。 • Atlas 300I Duo 推理卡：不支持。	• FP16：支持 • BF16：不支持	不支持	77	链接 链接 链接
Stable Diffusion XL_inpainting	• Atlas 800I A2 推理产品（32G）：支持的卡数为1。 • Atlas 300I Duo 推理卡：不支持。	• FP16：支持 • BF16：不支持	不支持	77	链接
Stable Diffusion XL_prompt_weight	• Atlas 800I A2 推理产品（32G）：支持的卡数为1。 • Atlas 300I Duo 推理卡：支持的卡数为1。	• FP16：支持 • BF16：不支持	不支持	77	链接
Stable Diffusion 3	• Atlas 800I A2 推理产品（32G）：支持的卡数为1或2。 • Atlas 300I Duo 推理卡：支持的卡数为1，可双芯并行。	• FP16：支持 • BF16：不支持	不支持	300	链接
Stable Video Diffusion	• Atlas 800I A2 推理产品（32G）：支持的卡数为1或2。 • Atlas 300I Duo 推理卡：不支持。	• FP16：支持 • BF16：不支持	不支持	不涉及	链接
Stable Audio Open v1.0	• Atlas 800I A2 推理产品（32G）：支持的卡数为1。 • Atlas 300I Duo 推理卡：支持的卡数为1。	• FP16：支持 • BF16：不支持	不支持	300	链接

3 使用场景

MindIE使用场景分为大模型服务化部署、大模型推理迁移流程和传统模型推理迁移流程三大场景，详情如下所示。

说明

对于大模型服务化部署场景，当前版本仅支持**2.1 大语言模型列表**中的模型。

场景	涉及组件	具体步骤	链接
大模型服务化部署	MindIE Service	1. 环境准备： a. 部署Kubernetes、MindCluster。（仅集群场景使用） b. 部署MindIE MS。（仅集群场景使用） c. 部署MindIE Server。 2. 启动服务。 3. 服务调用： a. 使用MindIE Server推理服务。 b. 使用MindIE Client发送请求（包括模型推理、请求管理和服务状态查询，用户调用接口即可实现与MindIE Server通信）。 c. 使用MindIE Benchmark工具测试推理性能和精度。 4. 性能调优。	1. 环境准备： a. 《MindIE Service开发指南》的“集群服务部署 > **环境准备**”章节 b. 《MindIE Service开发指南》的“快速开始 > **环境准备**”章节 2. 《MindIE Service开发指南》的“快速开始 > **启动服务**”章节 3. 服务调用： a. 《MindIE Service开发指南》的“MindIE Service组件 > MindIE Server > **使用指导**”章节 b. 《MindIE Service开发指南》的“MindIE Service组件 > MindIE Client > **功能介绍**”章节 c. 《MindIE Service开发指南》的“MindIE Service组件 > MindIE Service Tools > MindIE Benchmark > **功能介绍**”章节 4. 《MindIE Service开发指南》的“**性能调优**”章节

场景	涉及组件	具体步骤	链接
	vLLM	1. 环境准备： a. 版本配套关系： ● 当前基于CANN包8.0.0，Python 3.10，torch 2.1.0进行环境部署与运行，请安装相关适配版本。 ● 当前适配vLLM版本包括v0.3.3和v0.4.2。 b. 安装HDK、CANN、PTA、MindIE配套版本。 c. 安装vllm和vllm-npu。 2. 服务端使用拉起服务脚本拉起vllm在线推理服务。 3. 客户端使用curl、requests等方式向服务端发送推理请求。	《MindIE开源第三方服务化框架适配开发指南》的“vLLM >**环境准备**”章节
	Triton	1. 环境准备： a. 版本配套关系： ● 当前MindIE_Backend基于CANN包 8.0.0，Python 3.10，torch 2.1.0进行环境部署与运行，请安装相关适配版本。 ● 当前MindIE_Backend适配Triton版本为r24.02。 b. 安装HDK、CANN、PTA、ATB Models。 c. 安装MindIE LLM。 d. 安装Triton Inference Server。 e. 安装Triton Client。 f. 安装MindIE Backend。 2. 启动服务执行推理。 3. Triton Client发送测试请求。	《MindIE开源第三方服务化框架适配开发指南》的“Triton > **环境准备**”章节

场景	涉及组件	具体步骤	链接
	TGI	1. 环境准备： a. MindIE部署，包括部署昇腾适配驱动和固件、CANN、Kernel、MindIE、PyTorch相关包、ATB Models模型库。 b. 安装Tgi-MindIE包，当前Tgi-MindIE适配TGI版本包括v0.9.4和v2.0.4。 2. 服务端使用拉起服务脚本拉起TGI在线推理服务。 3. 客户端使用curl、requests等方式向服务端发送推理请求。	《MindIE开源第三方服务化框架适配开发指南》的“TGI > **环境准备**”章节
	TEI	1. 环境准备： a. MindIE部署，包括部署昇腾适配驱动和固件、CANN、Kernel、MindIE、PyTorch相关包、ATB Models模型库。 b. 利用MindIE Torch对文本嵌入模型/重排序模型进行编译优化。 c. 根据教程进行代码依赖安装、编译。 2. 服务端使用拉起服务脚本拉起TEI在线推理服务。 3. 客户端使用curl、requests等方式向服务端发送推理请求。	《MindIE开源第三方服务化框架适配开发指南》的“TEI > **环境准备**”章节
大模型推理迁移流程	MindIE LLM	1. 配置MindIE LLM。 2. 获取模型、权重。 3. 权重转换。（可选） 4. 权重量化。（可选） 5. 推理。	1. 《MindIE安装指南》中“配置MindIE > **配置MindIE LLM**”章节 2. 《MindIE LLM开发指南》的“模型推理使用流程 > **ATB Models使用**”章节 3. 《MindIE LLM开发指南》的“特性介绍 > **量化特性介绍**”章节

场景	涉及组件	具体步骤	链接
	MindIE SD	1. 下载模型仓。 2. 配置环境变量。 3. 准备模型权重。 4. 文生视频推理： a. 导入依赖包。 b. 配置文生视频工作流。 c. 模型编译。 d. 模型推理。	《MindIE SD开发指南》的“**快速上手**”章节
传统模型推理迁移流程	• MindIE Torch • MindIE RT	1. 导入mindietorch框架。 2. 模型导出。 3. 模型编译。 4. 模型推理。 5. 资源释放。	《MindIE Torch开发指南》的“**模型迁移快速入门**”章节
	Onnx	1. 准备模型权重。 2. 执行推理： a. 安装依赖。 b. 导出onnx和om文件。 c. 实现前后处理代码。 d. 模型推理。	《CANN 开发工具指南》的“ATC工具 > **快速入门**”章节

4 模型开箱

4.1 快速介绍

本章节主要介绍如何在以下任一产品上快速开始使用MindIE大模型推理方案：

- Atlas 800I A2 推理服务器
- Atlas 800 推理服务器（型号：3000）配置Atlas 300I Duo 推理卡

以ChatGLM3-6B为例，具体包括模型推理和模型性能等内容。

4.2 环境准备

前提条件

物理机部署场景，需要在物理机安装NPU驱动固件以及部署Docker，执行如下步骤判断是否已安装NPU驱动固件和部署Docker。

- 执行以下命令查看NPU驱动固件是否安装。若出现类似如**图4-1**所示，说明已安装。否则请参见**表4-1**或**表4-2**进行安装。
  ```
  npu-smi info
  ```


图 4-1 回显信息

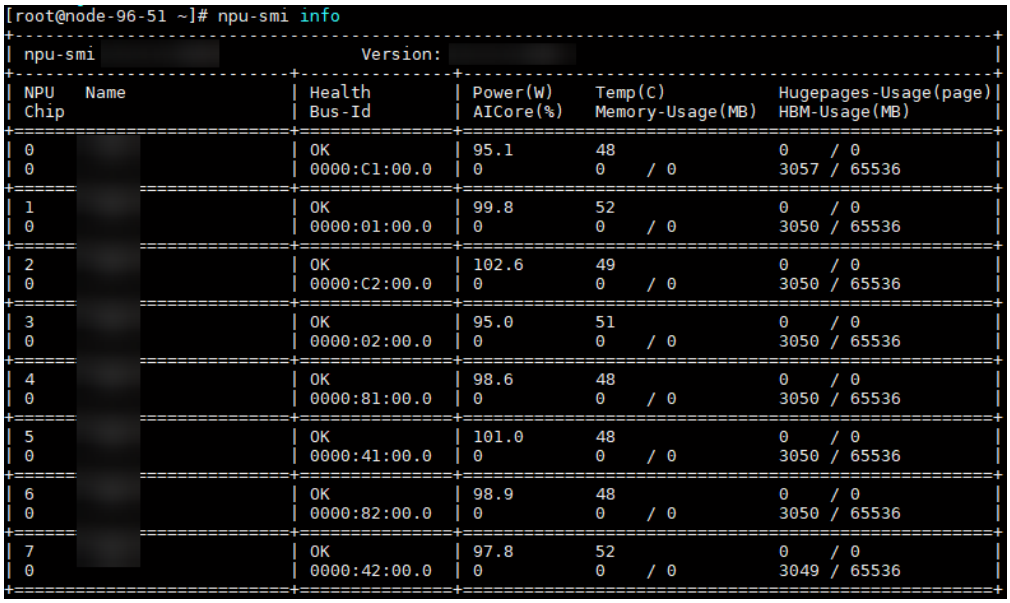


表 4-1 Atlas 推理系列产品

产品型号	参考文档
Atlas 300I Duo	《Atlas 中心推理卡 24.1.0 NPU驱动和固件安装指南》的“**物理机安装与卸载**”章节

表 4-2 Atlas 800I A2 推理产品

产品型号	参考文档
Atlas 800I A2	《Atlas A2 中心推理和训练硬件 24.1.0 NPU驱动和固件安装指南》中的“**物理机安装与卸载**”章节

- 执行以下命令查看Docker是否已安装并启动。

  ```
  docker ps
  ```

  回显以下信息表示Docker已安装并启动。

  ```
  CONTAINER ID  IMAGE  COMMAND  CREATED  STATUS  PORTS  NAMES
  ```

获取模型权重

请先下载权重，这里以ChatGLM3-6B为例，下载链接：**https://huggingface.co/THUDM/chatglm3-6b/tree/main**，将权重文件上传至服务器任意目录（如/home/weight）。

获取容器镜像

进入**昇腾官方镜像仓库**，根据设备型号选择下载对应的MindIE镜像。

该镜像已具备模型运行所需的基础环境，包括：CANN、FrameworkPTAdapter、MindIE与ATB Models，可实现模型快速上手推理。

表 4-3 容器内各组件安装路径

组件	安装路径
CANN	/usr/local/Ascend/ascend-toolkit
CANN-NNAL-ATB	/usr/local/Ascend/nnal/atb
MindIE	/usr/local/Ascend/mindie
ATB Models	/usr/local/Ascend/llm_model

4.3 框架选型

针对大语言模型推理，MindIE当前有三条推理框架组合方案的路线可供选择，三者都需要经过MindIE LLM的ATB Models模块，三条路线各有优劣势，具体可以根据用户使用场景选择。

图 4-2 路线框架图

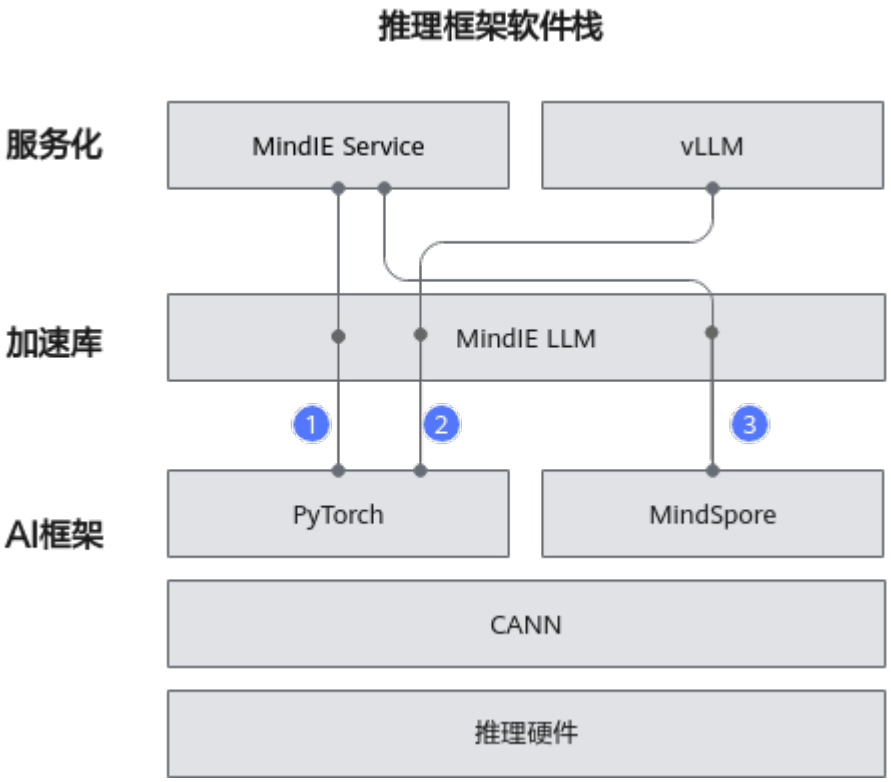


表 4-4 路线介绍说明

路线	典型场景	链接
路线1：MindIE Service + MindIE LLM + PyTorch	极致性能：在MindIE Service接口能够满足需求时，要求极致性能，推荐此路线。	请参见《MindIE Service开发指南》的“**产品简介**”章节。
路线2：vLLM+ MindIE LLM + PyTorch	接口丰富：用户AI应用，要求服务化接口功能齐全，推荐此路线。	请参见《MindIE开源第三方服务化框架适配开发指南》的“vLLM > **功能介绍**”章节。
路线3：MindIE Service + MindIE LLM + MindSpore	全栈可控：用户对于自主可控的诉求强烈，推荐此路线。	请参见《MindIE LLM开发指南》的“模型推理使用流程 > MindSpore Models使用 > **MindSpore Models服务化使用**”章节。

4.4 启动容器

步骤1 下载完成镜像后，执行以下命令启动容器。

```
docker run -it -d --net=host --shm-size=1g \
    --privileged \
    --name <container-name> \
    --device=/dev/davinci_manager \
    --device=/dev/hisi_hdc \
    --device=/dev/devmm_svm \
    -v /usr/local/Ascend/driver:/usr/local/Ascend/driver:ro \
    -v /usr/local/sbin:/usr/local/sbin:ro \
    -v /path-to-weights:/path-to-weights:ro \
    mindie:1.0.0-800I-A2-py311-openeuler24.03-lts bash
```

说明

“mindie:1.0.0-800I-A2-py311-openeuler24.03-lts”为镜像名称，可根据实际情况修改。

表 4-5 参数说明

参数	参数说明
--privileged	特权容器，允许容器访问宿主机的所有设备。
--name	设置容器名称。

参数	参数说明
--device	表示映射的设备，可以挂载一个或者多个设备。 需要挂载的设备如下： • /dev/davinci*X*：NPU设备，X是ID号，如：davinci0。 • /dev/davinci_manager：davinci相关的管理设备。 • /dev/hisi_hdc：hdc相关管理设备。 • /dev/devmm_svm：内存管理相关设备。 **说明** 可根据以下命令查询device个数及名称方式，根据需要绑定设备，修改上面命令中的"--device=****"。 `ll /dev/ | grep davinci`
-v /usr/local/Ascend/driver:/usr/local/Ascend/driver:ro	将宿主机目录“/usr/local/Ascend/driver”挂载到容器，请根据驱动所在实际路径修改。
-v /usr/local/sbin:/usr/local/sbin:ro	将宿主机工具“/usr/local/sbin/”以只读模式挂载到容器中，请根据实际情况修改。
-v /path-to-weights:/path-to-weights:ro	设定权重挂载的路径，需要根据用户的情况修改。

步骤2 执行以下命令进入容器。

```
docker exec -it <container-name> /bin/bash
```

----结束

说明

更多详细信息，请参考**启动容器**章节。

4.5 模型推理测试

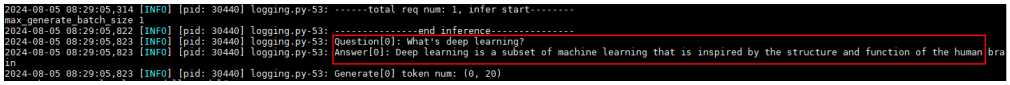
- 若为Atlas 800I A2 推理服务器请执行以下步骤。

  进入容器后，执行以下命令，运行ChatGLM3-6B推理样例。“/home/weight/chatglm3-6b”为容器的模型权重路径。

  ```
  source /usr/local/Ascend/ascend-toolkit/set_env.sh
  source /usr/local/Ascend/nnal/atb/set_env.sh
  source /usr/local/Ascend/atb-models/set_env.sh
  cd /usr/local/Ascend/llm_model
  bash examples/models/chatglm/v2_6b/run_800i_a2_pa.sh /home/weight/chatglm3-6b -
  trust_remote_code
  ```

  回显如**图4-3**所示信息，表示推理成功。

图 4-3 推理样例



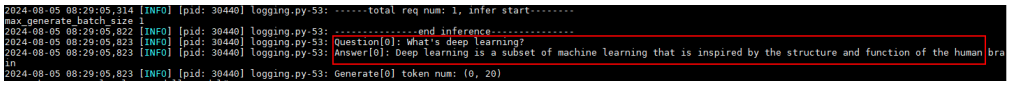
- 若为Atlas 800 推理服务器（型号：3000）配置Atlas 300I Duo 推理卡请执行以下步骤。

  进入容器后，执行以下命令，运行ChatGLM3-6B推理样例。“/home/weight/chatglm3-6b”为容器的模型权重路径。

  ```
  source /usr/local/Ascend/ascend-toolkit/set_env.sh
  source /usr/local/Ascend/nnal/atb/set_env.sh
  source /usr/local/Ascend/atb-models/set_env.sh
  cd /usr/local/Ascend/llm_model
  bash examples/models/chatglm/v2_6b/run_300i_duo_pa.sh /home/weight/chatglm3-6b -
  trust_remote_code
  ```

  回显如图4-4所示信息，表示推理成功。

图 4-4 推理样例



4.6 模型性能测试

说明

以ChatGLM3-6B为例，查看2.1 大语言模型列表该模型支持的卡数，使用命令“npu-smi info”可查看当前昇腾卡的占用情况。

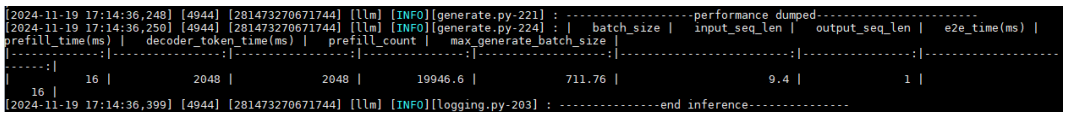
- 若为Atlas 800I A2 推理服务器请执行以下步骤。

步骤1 进入容器后，执行以下命令，运行ChatGLM3-6B模型性能测试。“/home/weight/chatglm3-6b”为容器的模型权重路径。

```
source /usr/local/Ascend/ascend-toolkit/set_env.sh
source /usr/local/Ascend/nnal/atb/set_env.sh
source /usr/local/Ascend/atb-models/set_env.sh
cd /usr/local/Ascend/llm_model/tests/modeltest
bash run.sh pa_fp16 performance [[2048,2048]] 16 chatglm /home/weight/chatglm3-6b 8 -
trust_remote_code
```

步骤2 模型性能测试时间大概2分钟，结束后回显如图4-5所示，代表模型性能测试成功。

图 4-5 性能测试样例



----结束

- 若为Atlas 800 推理服务器（型号：3000）配置Atlas 300I Duo 推理卡请执行以下步骤。

步骤1 进入容器后，执行如下命令，运行ChatGLM3-6B模型性能测试。“/home/weight/chatglm3-6b”为容器的模型权重路径。

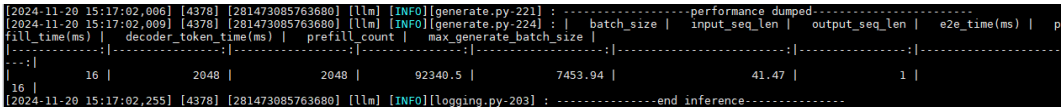
```
source /usr/local/Ascend/ascend-toolkit/set_env.sh
source /usr/local/Ascend/nnal/atb/set_env.sh
source /usr/local/Ascend/atb-models/set_env.sh
```



```
cd /usr/local/Ascend/llm_model/tests/modeltest
bash run.sh pa_fp16 performance [[2048,2048]] 16 chatglm /home/weight/chatglm3-6b 4 -
trust_remote_code
```

步骤2 模型性能测试时间大概5分钟，结束后回显如图4-6所示，代表模型性能测试成功。

图 4-6 性能测试样例



----结束

