AX NPU SDK 使用说明

文档版本：V1.4

发布日期：2022/03/14

AXERA CONFIDENTIAL FOR Sipeed

目 录

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

权利声明

爱芯元智半导体(上海)有限公司或其许可人保留一切权利。

非经权利人书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

注意

您购买的产品、服务或特性等应受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。除非商业合同另有约定，本公司对本文档内容不做任何明示或默示的声明或保证。

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。

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

前 言

适用产品

爱芯 AX620A

适读人群

- 软件开发工程师
- 技术支持工程师

符号与格式定义

符号/格式	说明
`xxx`	表示您可以执行的命令行。
斜体	表示变量。如，“*安装目录*/AX620_SDK_Vx.x.x/build 目录”中的“安装目录”是一个变量，由您的实际环境决定。
☞ 说明/备注：	表示您在使用产品的过程中，我们向您说明的事项。
！ 注意：	表示您在使用产品的过程中，需要您特别注意的事项。

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

修订历史

文档版本	发布时间	修订说明
V1.0	2021/08/24	文档初版
V1.1	2021/09/29	更新错误码
V1.2	2021/12/15	增加新接口函数 AX_NPU_SDK_EX_Init_with_attr_v1
V1.3	2022/01/24	更新函数说明
V1.4	2022/03/11	增加配置 NPU DDR 带宽

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

1 概述

AX NPU SDK 是在 Axera 平台上使用 NPU 执行推理任务的库。

- NPU 运行所需的输入输出内存均由用户提供。
- SDK 通过缓存结构体指针数组传递输入输出。
- 输入输出在缓存结构体指针数组中的顺序必须与 IO INFO 所描述的 tensor meta 顺序一致。
- SDK 提供通过 tensor meta 申请整块输入输出内存的接口和相应的释放接口。

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

2 功能

2.1 初始化

使用 SDK 时首先应进行初始化，初始化会分配 SDK 内部所用的各种资源。

2.2 反初始化

SDK 使用完毕时应进行反初始化，反初始化会释放 SDK 持有的所有资源。

2.3 创建模型句柄

通过 dot-neu 模型文件可以创建模型句柄，用于获取模型信息，执行推理任务等接口。

2.4 销毁模型句柄

模型使用完毕后应销毁模型句柄，释放与此模型相关的所有资源。

2.5 执行推理

通过模型句柄，输入数据和输出缓存进行推理，SDK 了提供同步推理接口和异步推理接口。

2.6 取消任务

已发起的异步推理任务可以通过此接口取消。

2.7 重置 NPU

NPU 出现硬件错误时可以通过此接口使 NPU 恢复到初始状态。

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

2.8 申请输入输出缓存

用户可以通过此接口一次性完成一块输入缓存或一块输出缓存的申请。

2.9 释放输入输出缓存

对于用户使用 SDK“申请输入输出缓存”接口申请的缓存，须使用此接口进行释放。

2.10 获取模型信息

通过模型句柄获取模型各个输入和输出的信息。

2.11 配置 NPU DDR 带宽

在调用 AX_NPU_SDK_EX_Init()初始化 NPU 之后，可以通过文件 npu_ddr_bw_limit_ctrl.sh 来配置 NPU DDR 带宽限制，具体使用方式可以参考 npu_ddr_bw_limit_ctrl.sh 文件中的 usage 信息，该文件位置在系统文件目录/soc/scripts/下。

该脚本和 AX_NPU_SDK_EX_Init_with_attr_v1()函数具有相同配置 NPU DDR 带宽限制功能，建议实际使用只选择其中一种方式。

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

2. 调用流程

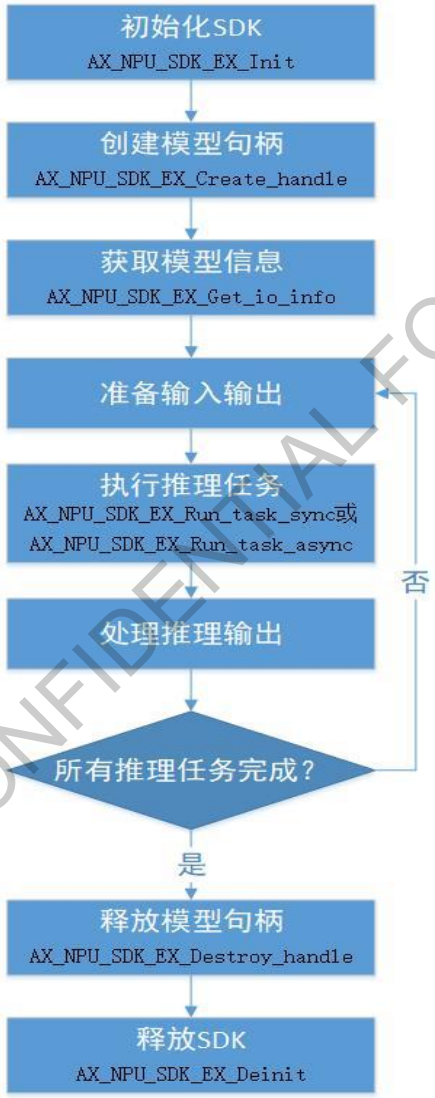

图3-1 调用流程

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

4 API 参考

AX_NPU_SDK_EX_Init

【描述】

初始化 NPU SDK。

【语法】

```
AX_S32 AX_NPU_SDK_EX_Init(AX_VOID);
```

【参数】

无

【返回值】

返回值	描述
0	成功
非 0	失败，返回错误码

【需求】

- 头文件：ax_interpreter_external_api.h
- 库文件：libax_interpreter_external.so

【注意】

不可以重复调用，调用前可以通过 AX_NPU_SDK_EX_Get_Attr 接口来获取是否已经初始化，以及初始化类型。

【举例】

无

【相关主题】

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_Deinit

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_Init_with_attr

【描述】

初始化 NPU SDK。

【语法】

```
AX_S32 AX_NPU_SDK_EX_Init_with_attr(AX_NPU_SDK_EX_ATTR_T *pNpuAttr);
```

【参数】

参数名称	描述	输入/输出
AX_NPU_SDK_EX_ATTR_T *pNpuAttr	NPU 运行模式	输入

【返回值】

返回值	描述
0	成功
非 0	失败，返回错误码

【需求】

- 头文件：ax_interpreter_external_api.h
- 库文件：libax_interpreter_external.so

【注意】

不可以重复调用，调用前可以通过 AX_NPU_SDK_EX_Get_Attr 接口来获取是否已经初始化，以及初始化类型。

【举例】

无

【相关主题】

AX_NPU_SDK_EX_Deinit

AVERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_Init_with_attr_v1

【描述】

初始化 NPU SDK，输入参数包含 NPU 运行模式和 NPU DDR 带宽限制

【语法】

```
AX_S32 AX_NPU_SDK_EX_Init_with_attr_v1(AX_NPU_SDK_EX_ATTR_V1_T *pNpuAttr_v1);
```

【参数】

参数名称	描述	输入/输出
AX_NPU_SDK_EX_ATTR_V1_T *pNpuAttr	NPU 运行模式 NPU 带宽限制	输入

【返回值】

返回值	描述
0	成功
非 0	失败，返回错误码

【需求】

- 头文件：ax_interpreter_external_api.h
- 库文件：libax_interpreter_external.so

【注意】

不可以重复调用，调用前可以通过 AX_NPU_SDK_EX_Get_Attr 接口来获取是否已经初始化，以及初始化类型。

【举例】

无

【相关主题】

AX_NPU_SDK_EX_Deinit

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_Deinit

【描述】

释放 IO 和内存和模型，退出线程。

【语法】

```
AX_VOID AX_NPU_SDK_EX_Deinit(AX_VOID);
```

【参数】

无

【返回值】

返回值	描述
无	无

【需求】

- 头文件：ax_interpreter_external_api.h
- 库文件：libax_interpreter_external.so

【注意】

无

【举例】

无

【相关主题】

AX_NPU_SDK_EX_Init

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_Create_handle

【描述】

加载 dot-neu，创建模型句柄，dot-neu 文件的大小就是模型加载后会占用的内存空间大小。

【语法】

```
AX_S32 AX_NPU_SDK_EX_Create_handle(AX_NPU_SDK_EX_HANDLE_T *handle, const
AX_VOID *dotNeuAddr, AX_S32 dotNeuLen);
```

【参数】

参数名称	描述	输入/输出
AX_NPU_SDK_EX_HANDLE_T *handle	模型句柄	输出
const AX_VOID *dotNeuAddr	模型数据起始地址	输入
AX_S32 dotNeuLen	模型数据长度	输入

【返回值】

返回值	描述
0	成功
非 0	失败，返回错误码

【需求】

- 头文件：ax_interpreter_external_api.h
- 库文件：libax_interpreter_external.so

【注意】

最多可以同时创建 1000 个模型句柄。

【举例】

无

【相关主题】

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_Destroy_handle

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_Destroy_handle

【描述】

释放模型句柄。

【语法】

```
AX_S32 AX_NPU_SDK_EX_Destroy_handle(AX_NPU_SDK_EX_HANDLE_T handle);
```

【参数】

参数名称	描述	输入/输出
AX_NPU_SDK_EX_HANDLE_T *handle	模型句柄	输入

【返回值】

返回值	描述
0	成功
非 0	失败，返回错误码

【需求】

- 头文件：ax_interpreter_external_api.h
- 库文件：libax_interpreter_external.so

【注意】

无

【举例】

无

【相关主题】

AX_NPU_SDK_EX_Create_handle

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_Run_task_sync

【描述】

模型同步推理。

【语法】

```
AX_S32 AX_NPU_SDK_EX_Run_task_sync(AX_NPU_SDK_EX_HANDLE_T handle,
AX_NPU_SDK_EX_IO_T *io);
```

【参数】

参数名称	描述	输入/输出
AX_NPU_SDK_EX_HANDLE_T *handle	模型句柄	输入
AX_NPU_SDK_EX_IO_T* io	模型输入输出	输入/输出

【返回值】

返回值	描述
0	成功
非 0	失败，返回错误码

【需求】

- 头文件：ax_interpreter_external_api.h
- 库文件：libax_interpreter_external.so

【注意】

无

【举例】

无

【相关主题】

无

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_Run_task_sync_v2

【描述】

模型同步推理。

【语法】

```
AX_S32 AX_NPU_SDK_EX_Run_task_sync_v2(AX_NPU_SDK_EX_HANDLE_T handle,
AX_NPU_SDK_EX_PREPROCESS_MICRO_CODE_T preprocessMicroCode, AX_NPU_SDK_EX_IO_T
*io);
```

【参数】

参数名称	描述	输入/输出
AX_NPU_SDK_EX_HANDLE_T *handle	模型句柄	输入
AX_NPU_SDK_EX_PREPROCESS_MICRO_CODE_T preprocessMicroCode	mcode	输入
AX_NPU_SDK_EX_IO_T* io	模型输入输出	输入/输出

【返回值】

返回值	描述
0	成功
非 0	失败，返回错误码

【需求】

- 头文件：ax_interpreter_external_api.h
- 库文件：libax_interpreter_external.so

【注意】

无

【举例】

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

无

【相关主题】

无

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_Run_task_async

【描述】

模型异步推理。

【语法】

```
AX_S32 AX_NPU_SDK_EX_Run_task_async(AX_NPU_SDK_EX_HANDLE_T handle,
AX_NPU_SDK_EX_RESOURCE_T *resource, AX_NPU_SDK_EX_TASK_ID_T *taskId);
```

【参数】

参数名称	描述	输入/输出
AX_NPU_SDK_EX_HANDLE_T *handle	模型句柄	输入
AX_NPU_SDK_EX_RESOURCE_T *resource	异步推理资源，包括 io，callback 和 user data	输入
AX_NPU_SDK_EX_TASK_ID_T *taskId	任务 id，uint32_t	输出

【返回值】

返回值	描述
0	成功
非 0	失败，返回错误码

【需求】

- 头文件：ax_interpreter_external_api.h
- 库文件：libax_interpreter_external.so

【注意】

无

【举例】

无

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_Run_task_async_v2

【描述】

模型异步推理。

【语法】

```
AX_S32 AX_NPU_SDK_EX_Run_task_async_v2(AX_NPU_SDK_EX_HANDLE_T handle,
AX_NPU_SDK_EX_PREPROCESS_MICRO_CODE_T preprocessMicroCode,
AX_NPU_SDK_EX_RESOURCE_T *resource, AX_NPU_SDK_EX_TASK_ID_T *taskId);
```

【参数】

参数名称	描述	输入/输出
AX_NPU_SDK_EX_HANDLE_T *handle	模型句柄	输入
AX_NPU_SDK_EX_PREPROCESS_MICRO_CODE_T preprocessMicroCode	mcode	输入
AX_NPU_SDK_EX_RESOURCE_T *resource	异步推理资源，包括 io，callback 和 user data	输入
AX_NPU_SDK_EX_TASK_ID_T *taskId	任务 id，uint32_t	输出

【返回值】

返回值	描述
0	成功
非 0	失败，返回错误码

【需求】

- 头文件：ax_interpreter_external_api.h
- 库文件：libax_interpreter_external.so

【注意】

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

无

【举例】

无

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_Cancel_task

【描述】

取消异步的推理任务。

【语法】

```
AX_S32 AX_NPU_SDK_EX_Cancel_task(AX_NPU_SDK_EX_TASK_ID_T taskId);
```

【参数】

参数名称	描述	输入/输出
AX_NPU_SDK_EX_TASK_ID_T taskId	异步推理任务 id	输入

【返回值】

返回值	描述
0	成功
非 0	失败，返回错误码

【需求】

- 头文件：ax_interpreter_external_api.h
- 库文件：libax_interpreter_external.so

【注意】

无

【举例】

无

【相关主题】

AX_NPU_SDK_EX_Run_task_async

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_Hard_reset

【描述】

重置 NPU。

【语法】

```
AX_VOID AX_NPU_SDK_EX_Hard_reset(AX_VOID);
```

【参数】

无

【返回值】

无

【需求】

- 头文件：ax_interpreter_external_api.h
- 库文件：libax_interpreter_external.so

【注意】

无

【举例】

无

【相关主题】

无

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_Alloc_buffer

【描述】

分配输入或输出的内存空间。

【语法】

```
AX_S32 AX_NPU_SDK_EX_Alloc_buffer(
    AX_NPU_SDK_EX_TENSOR_META_T* meta,
    AX_NPU_SDK_EX_BUF_T*  buf,
    AX_NPU_SDK_EX_ALLOC_BUFFER_STRATEGY_T strategy);
```

【参数】

参数名称	描述	输入/输出
AX_NPU_SDK_EX_TENSOR_META_T* meta	tensor meta, 包括 name，shape，bit 等信息	输入
AX_NPU_SDK_EX_BUF_T *buf	输入缓存或输出缓存	输出
AX_NPU_SDK_EX_ALLOC_BUFFER_STRATEGY_T strategy	内存分配策略参数	参数

【返回值】

返回值	描述
0	成功
非 0	失败，返回错误码

【需求】

- 头文件：ax_interpreter_external_api.h
- 库文件：libax_interpreter_external.so

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

【注意】

此函数将根据 meta 信息，分配物理和/或虚拟内存空间到 buf。使用时需要为用户的输入输出分别调用此函数分配输入输出 buffer。

【举例】

无

【相关主题】

AX_NPU_SDK_EX_Free_buffer

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_Free_buffer

【描述】

释放输入或输出的内存空间。

【语法】

```
AX_S32 AX_NPU_SDK_EX_Free_buffer(AX_NPU_SDK_EX_BUF_T* buf);
```

【参数】

参数名称	描述	输入/输出
AX_NPU_SDK_EX_BUF_T *buf	输入缓存或输出缓存	输入

【返回值】

返回值	描述
0	成功
非 0	失败，返回错误码

【需求】

- 头文件：ax_interpreter_external_api.h
- 库文件：libax_interpreter_external.so

【注意】

无

【举例】

无

【相关主题】

AX_NPU_SDK_EX_Alloc_buffer

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_Get_io_info

【描述】

获取模型的输入输出信息。

【语法】

```
const AX_NPU_SDK_EX_IO_INFO_T* AX_NPU_SDK_EX_Get_io_info
(AX_NPU_SDK_EX_HANDLE_T handle);
```

【参数】

参数名称	描述	输入/输出
AX_NPU_SDK_EX_HANDLE_T handle	模型句柄	输入

【返回值】

返回值	描述
const AX_NPU_SDK_EX_IO_INFO_T*	记录模型输入输出信息的结构体指针

【需求】

- 头文件：ax_interpreter_external_api.h
- 库文件：libax_interpreter_external.so

【注意】

无

【举例】

无

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_Get_Attr

【描述】

获取虚拟 NPU 运行属性。

【语法】

```
AX_S32 AX_NPU_SDK_EX_Get_Attr(AX_NPU_SDK_EX_ATTR_T *pNpuAttr)
```

【参数】

参数名称	描述	输入/输出
AX_NPU_SDK_EX_ATTR_T *pNpuAttr	虚拟 NPU 运行模式	输入/输出

【返回值】

返回值	描述
0	获取属性成功
非 0	获取属性失败

【需求】

- 头文件：ax_interpreter_external_api.h
- 库文件：libax_interpreter_external.so

【注意】

无

【举例】

无

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_Get_Model_type

【描述】

获取虚拟 NPU 运行模型类型。

【语法】

```
AX_S32 AX_NPU_SDK_EX_Get_Model_type(AX_NPU_SDK_EX_HANDLE_T handle,
AX_NPU_SDK_EX_MODEL_TYPE_T *pModelType)
```

【参数】

参数名称	描述	输入/输出
AX_NPU_SDK_EX_HANDLE_T handle	模型句柄	输入
AX_NPU_SDK_EX_MODEL_TYPE_T *pModelType	虚拟 NPU 模型类型	输入/输出

【返回值】

返回值	描述
0	获取属性成功
非 0	获取属性失败

【需求】

- 头文件：ax_interpreter_external_api.h
- 库文件：libax_interpreter_external.so

【注意】

无

【举例】

无

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_Get_Dot_neu_type

【描述】

获取虚拟 NPU 运行模型类型。

【语法】

```
AX_S32 AX_NPU_SDK_EX_Get_Dot_neu_type(const AX_VOID *dotNeuAddr, AX_S32
dotNeuLen, AX_NPU_SDK_EX_MODEL_TYPE_E *pModelType)
```

【参数】

参数名称	描述	输入/输出
const AX_VOID *dotNeuAddr	模型数据起始地址	输入
AX_S32 dotNeuLen	模型数据长度	输入
AX_NPU_SDK_EX_MODEL_TYPE_T *pModelType	虚拟 NPU 模型类型	输入/输出

【返回值】

返回值	描述
0	获取属性成功
非 0	获取属性失败

【需求】

- 头文件：ax_interpreter_external_api.h
- 库文件：libax_interpreter_external.so

【注意】

无

【举例】

无

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

5 数据结构

AX_NPU_SDK_EX_BUF_T

【说明】

输入或输出缓存，包括用户缓存和 SDK 内部使用的缓存。

【定义】

```
typedef struct {
    AX_ADDR phyAddr;
    AX_VOID *pVirAddr;
    AX_U32 nSize;
    AX_ADDR innerPhyAddr;
    AX_VOID *pInnerVirAddr;
    AX_U32 nInnerSize;
} AX_NPU_SDK_EX_BUF_T
```

【成员】

成员名称	描述
AX_ADDR phyAddr	用户缓存的物理地址
AX_VOID *pVirAddr	用户缓存的虚拟地址
AX_U32 nSize	用户缓存的长度
AX_ADDR innerPhyAddr	SDK 内部缓存的物理地址
AX_VOID *pInnerVirAddr	SDK 内部缓存时虚拟地址
AX_U32 nInnerSize	SDK 内部缓存的长度

【注意】

用户应使用 AX_NPU_SDK_EX_Alloc_buffer 创建输入输出缓存，此接口可以正确处理用户缓

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

存与 SDK 内部缓存的关系。使用完毕后需要使用 AX_NPU_SDK_EX_Free_buffer 释放。

【相关数据类型及接口】

结构体：

- AX_NPU_SDK_EX_IO_T
- AX_NPU_SDK_EX_Alloc_buffer()
- AX_NPU_SDK_EX_Free_buffer()

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_ATTR_T

【说明】

虚拟 NPU 运行模型。

【定义】

```
typedef struct {
    AX_NPU_SDK_EX_HARD_MODE_T eHardMode;
} AX_NPU_SDK_EX_ATTR_T;
```

【成员】

成员名称	描述
AX_NPU_SDK_EX_HARD_MODE_T eHardMode	输入的数组

【相关数据类型及接口】

```
typedef enum {
    AX_NPU_VIRTUAL_DISABLE = 0,
    AX_NPU_VIRTUAL_3_1 = 1,
    AX_NPU_VIRTUAL_2_2 = 2,
    AX_NPU_VIRTUAL_1_1 = 3
} AX_NPU_SDK_EX_HARD_MODE_T;
```

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_ATTR_V1_T

【说明】

虚拟 NPU 运行模型和 DDR 带宽限制

【定义】

```
typedef struct {
    AX_NPU_SDK_EX_HARD_MODE_T eHardMode;
    AX_NPU_SDK_EX_DDR_LIM_T eDdrBwLimit
} AX_NPU_SDK_EX_ATTR_T;
```

【成员】

成员名称	描述
AX_NPU_SDK_EX_HARD_MODE_T eHardMode	输入的数组
AX_NPU_SDK_EX_DDR_LIM_T eDdrBwLimit	输入参数

【相关数据类型及接口】

```
typedef enum {
    AX_NPU_VIRTUAL_DISABLE = 0,
    AX_NPU_VIRTUAL_3_1 = 1,
    AX_NPU_VIRTUAL_2_2 = 2,
    AX_NPU_VIRTUAL_1_1 = 3
} AX_NPU_SDK_EX_HARD_MODE_T;
typedef enum {
    AX_NPU_EX_DDR_BW_LIMIT_DISABLE = 0,
    AX_NPU_EX_DDR_BW_LIMIT_2P1GB = 1,
    AX_NPU_EX_DDR_BW_LIMIT_2P8GB = 2,
    AX_NPU_EX_DDR_BW_LIMIT_3P5GB = 3,
```

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

```
    AX_NPU_EX_DDR_BW_LIMIT_4P1GB = 4,
    AX_NPU_EX_DDR_BW_LIMIT_4P8GB = 5,
    AX_NPU_EX_DDR_BW_LIMIT_5P5GB = 6,
    AX_NPU_EX_DDR_BW_LIMIT_6P2GB = 7,
    AX_NPU_EX_DDR_BW_LIMIT_6P9GB = 8,
    AX_NPU_EX_DDR_BW_LIMIT_7P6GB = 9,
    AX_NPU_EX_DDR_BW_LIMIT_8P3GB = 10,
    AX_NPU_EX_DDR_BW_LIMIT_9P0GB = 11,
    AX_NPU_EX_DDR_BW_LIMIT_9P9GB = 12,
    AX_NPU_EX_DDR_BW_LIMIT_MAX
} AX_NPU_SDK_EX_DDR_LIM_T;
```

AX_NPU_SDK_EX_IO_T

【说明】

模型所有输入输出。

【定义】

```
typedef struct {
    AX_NPU_SDK_EX_BUF_T *pInputs;
    AX_U32 nInputSize;
    AX_NPU_SDK_EX_BUF_T *pOutputs;
    AX_U32 nOutputSize;
} AX_NPU_SDK_EX_IO_T
```

【成员】

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

成员名称	描述
AX_NPU_SDK_EX_BUF_T *pInputs	输入的数组
AX_U32 nInputSize	输入的数量
AX_NPU_SDK_EX_BUF_T *pOutputs	输出的数组
AX_U32 nOutputSize	输出的数量

【注意】

- 输入数组 pInputs 中各个输入必须与 AX_NPU_SDK_EX_Get_io_info 接口返回的结构体中 pInputs 数组所表示的顺序完全一致；
- 输出数组 pOutputs 中各个输入必须与 AX_NPU_SDK_EX_Get_io_info 接口返回的结构体中 pOutputs 数组所表示的顺序完全一致。

【相关数据类型及接口】

结构体：

- AX_NPU_SDK_EX_RESOURCE_T
- AX_S32 AX_NPU_SDK_EX_Run_task_sync()

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_RESOURCE_T

【说明】

包装异步推理所需的各项参数。

【定义】

```
typedef struct {
    AX_NPU_SDK_EX_IO_T tIo;
    AX_NPU_SDK_EX_FINISH_FUNC fnFinishFunc;
    AX_VOID *pUserData;
} AX_NPU_SDK_EX_RESOURCE_T;
```

【成员】

成员名称	描述
AX_NPU_SDK_EX_IO_T tIo	输入输出
AX_NPU_SDK_EX_FINISH_FUNC fnFinishFunc	回调函数
AX_VOID *pUserData	用户自定义数据

【注意】

无

【相关数据类型及接口】

AX_S32 AX_NPU_SDK_EX_Run_task_async()

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_MEMORY_TYPE_T

【说明】

内存类型。

【定义】

```
typedef enum {
    AX_NPU_MT_INVALID = 0,
    AX_NPU_MT_PHYSICAL = 1,
    AX_NPU_MT_VIRTUAL = 2
} AX_NPU_SDK_EX_MEMORY_TYPE_T;
```

【成员】

成员名称	描述
AX_NPU_MT_INVALID	非法内存类型
AX_NPU_MT_PHYSICAL	物理内存
AX_NPU_MT_VIRTUAL	虚拟内存

【注意】

- AX_NPU_MT_PHYSICAL 类型的内存使用 AX_SYS_API 进行申请和释放，比如 AX_SYS_MemAlloc/AX_SYS_MemFree
- AX_NPU_MT_VIRTUAL 类型的内存使用系统内存管理接口进行申请和释放，比如 malloc/free，new/delete

【相关数据类型及接口】

结构体：AX_NPU_SDK_EX_TENSOR_META_T

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_TENSOR_META_T

【说明】

输入输出的基本信息。

【定义】

```
typedef struct {
    AX_S8 *pName;
    AX_U32 *pShape;
    AX_U8 nShapeNDim;
    AX_U32 nBit;
    AX_U32 nInnerBit;
    AX_U32 nSize;
    AX_U32 nInnerSize;
    AX_NPU_SDK_EX_MEMORY_TYPE_T eMemoryType;
    AX_S32 nPreallocOCMStartAddr;
    AX_S32 nPreallocOCMEndAddr;
} AX_NPU_SDK_EX_TENSOR_META_T
```

【成员】

成员名称	描述
AX_S8 *pName	输入输出的名称
AX_U32 *pShape	输入输出的维度
AX_U8 nShapeNDim	输入输出的维度数量
AX_U32 nBit	输入输出的位宽
AX_U32 nInnerBit	SDK 内部某阶段输出的位宽

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

成员名称	描述
AX_U32 nSize	输入输出的大小
AX_U32 nInnerSize	SDK 内部某阶段输出的长度
AX_NPU_SDK_EX_MEMORY_TYPE_T eMemoryType	非 Inner 缓存的内存类型
AX_S32 nPreallocOCMStartAddr	OCM 预分配空间起始地址
AX_S32 nPreallocOCMEndAddr	OCM 预分配空间结束地址

【注意】

- 由于某些模型的 NPU 直接输出还需在 CPU 上做一些计算才能得到最终的输出，所以推理时需要用户为 SDK 提供 Inner。
- Inner 的内存类型必定为 AX_NPU_MT_PHYSICAL。当 eMemoryType 为 AX_NPU_MT_PHYSICAL 时，无须为非 Inner 额外分配内存，可以与 Inner 共用同一块内存（此时 nSize 与 nInnerSize 必定相等）；当 eMemoryType 为 AX_NPU_MT_VIRTUAL 时，需要为 Inner 分配 nInnerSize 大小的 AX_NPU_MT_PHYSICAL 类型内存，为非 Inner 分配 nSize 大小的 AX_NPU_MT_VIRTUAL 类型内存。
- SDK 提供的 AX_NPU_SDK_EX_Alloc_buffer 接口可以正确处理 Inner 与非 Inner 的内存分配，建议使用此接口。

【相关数据类型及接口】

结构体：

- AX_NPU_SDK_EX_IO_INFO_T
- AX_S32 AX_NPU_SDK_EX_Alloc_buffer()

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_IO_INFO_T

【说明】

模式所有输入输出信息。

【定义】

```
typedef struct {
    AX_NPU_SDK_EX_TENSOR_META_T *pInputs;
    AX_U32 nInputSize;
    AX_NPU_SDK_EX_TENSOR_META_T *pOutputs;
    AX_U32 nOutputSize;
} AX_NPU_SDK_EX_IO_INFO_T
```

【成员】

成员名称	描述
AX_NPU_SDK_EX_TENSOR_META_T *pInputs	输入信息的数组
AX_U32 nInputSize	输入的数量
AX_NPU_SDK_EX_TENSOR_META_T *pOutputs	输出信息的数组
AX_U32 nOutputSize	输出的数量

【注意】

无

【相关数据类型及接口】

结构体：

- AX_VIN_DEV_T
- const AX_NPU_SDK_EX_IO_INFO_T* AX_NPU_SDK_EX_Get_io_info()

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_ALLOC_BUFFER_STRATEGY_T

【说明】

内存分配策略的枚举。

【定义】

```
typedef enum {

    AX_NPU_ABST_DEFAULT = 0,

} AX_NPU_SDK_EX_ALLOC_BUFFER_STRATEGY_T;
```

【成员】

成员名称	描述
AX_NPU_ABST_DEFAULT	默认内存分配策略（不带 cache）

【注意】

无

【相关数据类型及接口】

AX_NPU_SDK_EX_Alloc_buffer()

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_HARD_MODE_T

【说明】

虚拟 NPU 运行模式。

【定义】

```
typedef enum {
    AX_NPU_VIRTUAL_DISABLE = 0,
    AX_NPU_VIRTUAL_3_1 = 1,
    AX_NPU_VIRTUAL_2_2 = 2,
    AX_NPU_VIRTUAL_1_1 = 3
} AX_NPU_SDK_EX_HARD_MODE_T;
```

【成员】

成员名称	描述
AX_NPU_VIRTUAL_DISABLE = 0	禁用虚拟 NPU
AX_NPU_VIRTUAL_3_1	3/1 分配虚拟 NPU
AX_NPU_VIRTUAL_2_2	2/2 分配虚拟 NPU
AX_NPU_VIRTUAL_1_1	1/1 分配虚拟 NPU

【注意】

```
AX_NPU_VIRTUAL_3_1 = 1,
AX_NPU_VIRTUAL_2_2 = 2,
```

以上 2 项 AX620 上不支持

【相关数据类型及接口】

无

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_MODEL_TYPE_T

【说明】

虚拟 NPU 模型类型。

【定义】

```
typedef enum {
    AX_NPU_MODEL_TYPE_DEFUALT = 0,
      AX_NPU_MODEL_TYPE_3_1_1 = 1,
      AX_NPU_MODEL_TYPE_3_1_2 = 2,
      AX_NPU_MODEL_TYPE_2_2_1 = 3,
      AX_NPU_MODEL_TYPE_2_2_2 = 4,
      AX_NPU_MODEL_TYPE_1_1_1 = 5,
      AX_NPU_MODEL_TYPE_1_1_2 = 6,
} AX_NPU_SDK_EX_MODEL_TYPE_T;
```

【成员】

成员名称	描述
AX_NPU_MODEL_TYPE_DEFUALT	单 NPU 模型
AX_NPU_MODEL_TYPE_3_1_1	3/1 分配虚拟 NPU 模型中的第 1 个模型
AX_NPU_MODEL_TYPE_3_1_2	3/1 分配虚拟 NPU 模型中的第 2 个模型
AX_NPU_MODEL_TYPE_2_2_1	2/2 分配虚拟 NPU 模型中的第 1 个模型
AX_NPU_MODEL_TYPE_2_2_2	2/2 分配虚拟 NPU 模型中的第 2 个模型
AX_NPU_MODEL_TYPE_1_1_2	1/1 分配虚拟 NPU 模型中的第 1 个模型
AX_NPU_MODEL_TYPE_1_1_2	1/1 分配虚拟 NPU 模型中的第 2 个模型

【注意】

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

```
AX_NPU_MODEL_TYPE_3_1_1 = 1,

AX_NPU_MODEL_TYPE_3_1_2 = 2,

AX_NPU_MODEL_TYPE_2_2_1 = 3,

AX_NPU_MODEL_TYPE_2_2_2 = 4,
```

以上在 AX620 上不支持

【相关数据类型及接口】

无

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

AX_NPU_SDK_EX_DDR_LIM_T

【说明】

NPU DDR 带宽控制选项。

【定义】

```
typedef enum {
    AX_NPU_EX_DDR_BW_LIMIT_DISABLE = 0,
    AX_NPU_EX_DDR_BW_LIMIT_2P1GB = 1,
    AX_NPU_EX_DDR_BW_LIMIT_2P8GB = 2,
    AX_NPU_EX_DDR_BW_LIMIT_3P5GB = 3,
    AX_NPU_EX_DDR_BW_LIMIT_4P1GB = 4,
    AX_NPU_EX_DDR_BW_LIMIT_4P8GB = 5,
    AX_NPU_EX_DDR_BW_LIMIT_5P5GB = 6,
    AX_NPU_EX_DDR_BW_LIMIT_6P2GB = 7,
    AX_NPU_EX_DDR_BW_LIMIT_6P9GB = 8,
    AX_NPU_EX_DDR_BW_LIMIT_7P6GB = 9,
    AX_NPU_EX_DDR_BW_LIMIT_8P3GB = 10,
    AX_NPU_EX_DDR_BW_LIMIT_9P0GB = 11,
    AX_NPU_EX_DDR_BW_LIMIT_9P9GB = 12,
    AX_NPU_EX_DDR_BW_LIMIT_MAX
} AX_NPU_SDK_EX_DDR_LIM_T;
```

【成员】

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

成员名称	描述
AX_NPU_EX_DDR_BW_LIMIT_DISABLE	禁止带宽限制功能
AX_NPU_EX_DDR_BW_LIMIT_2P1GB	带宽限制最大值 2.1GB
AX_NPU_EX_DDR_BW_LIMIT_2P8GB	带宽限制最大值 2.8GB
AX_NPU_EX_DDR_BW_LIMIT_3P5GB	带宽限制最大值 3.5GB
AX_NPU_EX_DDR_BW_LIMIT_4P1GB	带宽限制最大值 4.1GB
AX_NPU_EX_DDR_BW_LIMIT_4P8GB	带宽限制最大值 4.8GB
AX_NPU_EX_DDR_BW_LIMIT_5P5GB	带宽限制最大值 5.5GB
AX_NPU_EX_DDR_BW_LIMIT_6P2GB	带宽限制最大值 6.2GB
AX_NPU_EX_DDR_BW_LIMIT_6P9GB	带宽限制最大值 6.9GB
AX_NPU_EX_DDR_BW_LIMIT_7P6GB	带宽限制最大值 7.6GB
AX_NPU_EX_DDR_BW_LIMIT_8P3GB	带宽限制最大值 8.3GB
AX_NPU_EX_DDR_BW_LIMIT_9P0GB	带宽限制最大值 9.0GB
AX_NPU_EX_DDR_BW_LIMIT_9P9GB	带宽限制最大值 9.9GB
AX_NPU_EX_DDR_BW_LIMIT_MAX	带宽选项最大值

【注意】

大于 AX_NPU_EX_DDR_BW_LIMIT_5P5GB（含） 的选项只用于调试目的。

【相关数据类型及接口】

无

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.

6 错误码

NPU SDK API 错误码如下表所示：

表6-1 NPU SDK API 错误码列表

错误代码	宏定义	描述
0x80060081	AX_NPU_DEV_STATUS_HANDLE_INVALID	无效句柄。
0x80060082	AX_NPU_DEV_STATUS_TASK_INVALID	无效任务。
0x80060083	AX_NPU_DEV_STATUS_NO_INIT	未初始化。
0x80060084	AX_NPU_DEV_STATUS_PARAM_INVALID	无效参数。
0x80060085	AX_NPU_DEV_STATUS_NO_RESOURCES	资源耗尽。
0x80060086	AX_NPU_DEV_STATUS_MEM_ERROR	内存错误。
0x80060087	AX_NPU_DEV_STATUS_HARD_ERROR	硬件错误。
0x80060088	AX_NPU_DEV_STATUS_NOT_SUPPORT	未支持。
0x80060089	AX_NPU_DEV_STATUS_TASK_BUSY	任务忙。
0x8006008a	AX_NPU_TASK_STATUS_FAILED	任务运行失败。

AXERA CONFIDENTIAL FOR Sipeed

The information contained in this Documentation may be privileged and confidential. If the reader of this information is not intended recipient, you are on notice that any distribution of this information, in any form, is strictly prohibited.