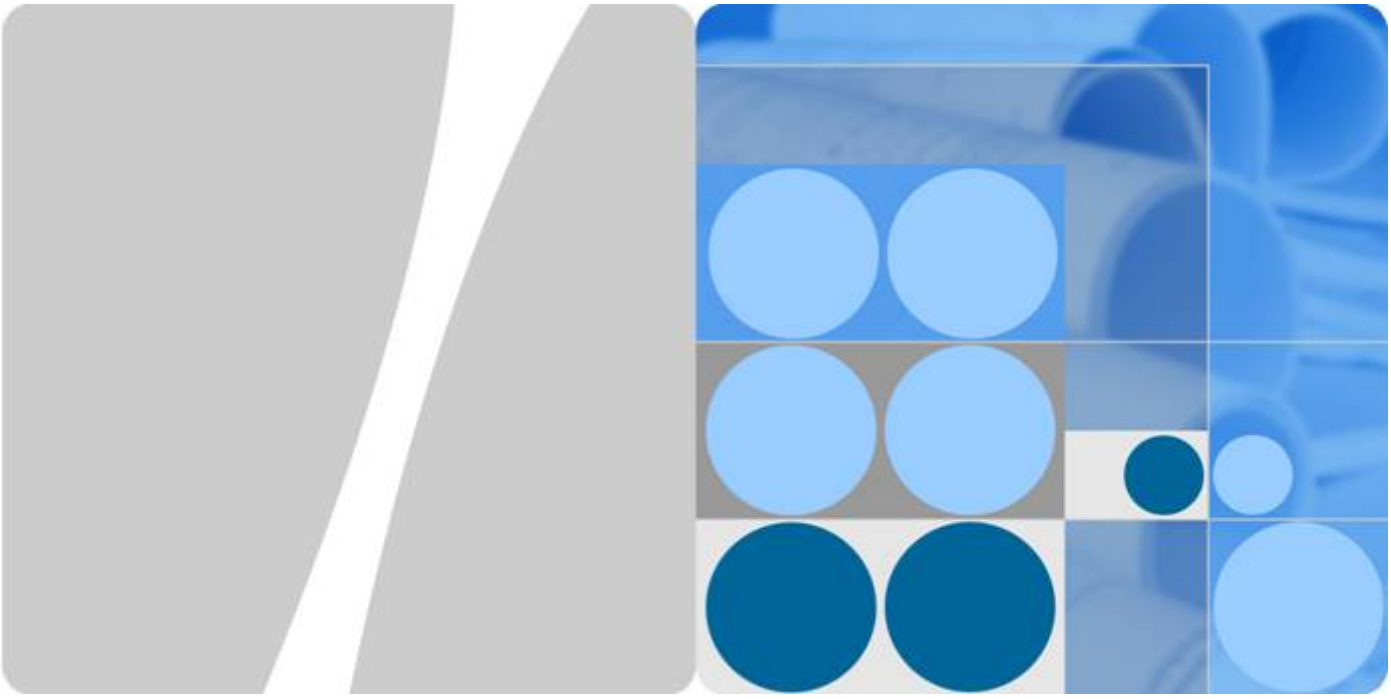


# 内容审核

# 产品介绍

**文档版本　01**

**发布日期　2024-07-01**

**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。

# 华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： https://www.huawei.com

客户服务邮箱： support@huawei.com

客户服务电话： 4008302118

# 目 录



# 1 什么是内容审核

内容审核（Content Moderation），是基于图像、文本、音频、视频、音频流的检测技术，可自动检测涉黄、图文违规等内容，对用户上传的图片、文字、音视频进行内容审核，以满足上传要求，帮助客户降低业务违规风险。

随着互联网的飞速发展和信息量猛增，大量色情等不良信息夹杂其中，如果不做好内容审核，不良内容会让用户产生反感，从而降低产品使用频率，最终远离产品。

内容审核以开放API（Application Programming Interface，应用程序编程接口）的方式提供给用户，用户通过调用API获取推理结果，帮助用户打造智能化业务系统，提升业务效率。

## 内容审核-图像

图像内容审核，利用深度神经网络模型对图片内容进行检测，准确识别图像中的涉黄内容等，帮助业务规避违规风险。

## 内容审核-文本

文本内容审核，采用人工智能文本检测技术有效识别涉黄、广告、辱骂和灌水文本内容，提供定制化的文本敏感内容审核方案。

**图 1-1** 文本内容审核示意图

## 内容审核-音频

基于领先的语音识别引擎、智能文本检测模型，精准识别出语音中涉黄、辱骂等违规场景，极大提升产品用户体验。

## 内容审核-视频

基于先进的人工智能技术综合检测视频画面、声音、字幕等，精准高效识别各类涉黄、涉暴、广告等违规内容，提高平台内容治理质量和效率。

## 内容审核-音频流

精准识别多场景下色情、辱骂、广告等违规内容，防御内容风险，提高音频流的审核效率，提升用户体验。

# 2 产品优势

## 检测准确

基于深度学习技术和大量的样本库，帮助客户快速准确进行违规内容检测，维护内容安全。

## 功能丰富

提供图文视频内容检测，覆盖涉黄、广告、涉暴等多种违规风险的内容检测，以及检测图像清晰度和构图质量等功能。

## 稳定可靠

内容审核服务已成功应用于各类场景，基于华为等企业客户的长期实践，经受过复杂场景考验。

## 简单高效

提供RESTful规范的API接口，以及服务SDK，方便客户使用与集成；帮助客户减少人力成本，节省业务支出。

# 3 应用场景

## 内容审核-图像

内容审核-图像有以下应用场景：

- 视频直播

  在互动直播场景中，成千上万个房间并发直播，人工审核直播内容几乎不可能。基于图像审核能力，可对所有房间内容实时监控，识别可疑房间并进行预警。

  场景优势如下：

  - 准确率高：基于改进的深度学习算法，检测准确率高。
  - 响应速度快：视频直播响应速度小于0.1秒。

- 在线商城

  智能审核商家/用户上传图像，高效识别并预警不合规图片，防止涉黄、涉暴类图像发布，降低人工审核成本和业务违规风险。

  场景优势如下：

  - 准确率高：基于改进的深度学习算法，检测准确率高。
  - 响应速度快：单张图像识别速度小于0.1秒。

- 网站论坛

  不合规图片的识别和处理是用户原创内容（UGC）类网站的重点工作，基于内容审核，可以识别并预警用户上传的不合规图片，帮助客户快速定位处理，降低业务违规风险。

  场景优势如下：

  - 准确率高：基于改进的深度学习算法，检测准确率高。
  - 响应速度快：单张图像识别速度小于0.1秒。

## 内容审核-文本

内容审核-文本有以下应用场景：

- 电商评论筛查

  审核电商网站产品评论，智能识别有色情、灌水等违规评论，保证良好用户体验。

  场景优势如下：

  - 准确率高：基于改进的深度学习算法，检测准确率高。
  - 响应速度快：响应速度小于0.1秒。
- 注册昵称审核

  对网站的用户注册信息进行智能审核，过滤包含广告、色情等内容的用户昵称。

  场景优势如下：
  - 准确率高：基于改进的深度学习算法，检测准确率高。
  - 响应速度快：响应速度小于0.1秒。
- 媒资内容审核

  自动识别媒资中可能存在的违禁品等信息，避免已发布的文章存在违规风险。

  场景优势如下：
  - 快速迭代：持续快速的迭代文本词库，及时识别新型不合规内容。
  - 处理速度快：处理速度小于0.1秒。
- 弹幕审核

  实时检测弹幕文本、保证网络直播间内容安全，降低业务违规风险。

  场景优势如下：
  - 海量词库：内置海量词库，支持各种匹配规则。
  - 快速迭代：持续快速的迭代文本词库，及时识别新型不合规内容。
- 聊天内容实时审核

  实时检测游戏等文本聊天内容中可能出现的违规信息，避免辱骂、色情、反动等文本内容，净化网络环境。

  场景优势如下：
  - 海量词库：内置海量词库，支持各种匹配规则。
  - 响应速度快：响应速度小于0.1秒。

## 内容审核-音频

内容审核-音频有以下应用场景：

- 在线教育

  监测在线教育中有声教学内容，智能审核音频中的涉黄、涉暴、辱骂、广告等违规场景。

  场景优势如下：
  - 准确率高：基于改进的深度学习算法，基于复杂环境语音审核准确率高。
  - 支持特殊声音识别：支持特殊声音识别模型，如娇喘、呻吟、敏感声纹等。
- 游戏/社交语音

  监测游戏APP / 社交APP中的聊天内容以及语音动态，降低业务违规风险。

  场景优势如下：
  - 准确率高：基于改进的深度学习算法，基于复杂环境语音审核准确率高。
  - 支持特殊声音识别：支持特殊声音识别模型，如娇喘、呻吟、敏感声纹等。
- 录播/电台语音

  监测内容传播类 / FM电台类音频数据，降低业务违规风险。

  场景优势如下：

- 准确率高：基于改进的深度学习算法，基于复杂环境语音审核准确率高。
- 支持特殊声音识别：支持特殊声音识别模型，如娇喘、呻吟、敏感声纹等。

## 内容审核-视频

内容审核-视频有以下应用场景：

- 视频平台/社区：精准识别平台上的违规视频内容，帮助平台规避内容风险：
  - 360度全方位检测：提供多模态综合审核方案，对视频内容中的画面、声音、文字进行全方位解析。
  - 支持类型广：支持多种视频文件格式：AVI、FLV、MP4、MPG、WMV、MOV、RMVB、M3U8等
- 视频聊天：精准识别和拦截社交/即时通讯场景下的色情、辱骂、暴恐、广告导流等违规内容：
  - 360度全方位检测：提供多模态综合审核方案，对视频内容中的画面、声音、文字进行全方位解析。
  - 支持类型广：支持多种视频文件格式：AVI、FLV、MP4、MPG、WMV、MOV、RMVB、M3U8等
- 在线教育：精准识别和拦截线上教学、互动、录播课程中的违规内容，保障用户尤其是未成年人的身心健康：
  - 360度全方位检测：提供多模态综合审核方案，对视频内容中的画面、声音、文字进行全方位解析。
  - 支持类型广：支持多种视频文件格式：AVI、FLV、MP4、MPG、WMV、MOV、RMVB、M3U8等。

## 内容审核-音频流

- 语音直播间

  语音直播间通过语音进行实时交流和互动，把音频流审核集成到语音直播平台以实现实时审核功能，实时判断出不合规的语音内容。

  场景优势：
  - 实时性：可以实时监测和分析直播间中的语音内容，保障直播间的秩序和安全。
  - 支持特殊声音识别：支持特殊声音识别模型，如娇喘、呻吟、敏感声纹等。
- 社交语音消息

  在社交语音消息平台上实时对用户发送的语音消息进行审核，及时判断出包含不良内容的语音消息，帮助您根据审核结果进行相应的处理，如删除消息、禁言用户等。

  场景优势：
  - 准确率高：全面场景覆盖，避免误杀漏杀，实时防御风险。
  - 支持特殊声音识别：支持特殊声音识别模型，如娇喘、呻吟、敏感声纹等。
- 在线教育

  根据教育内容和要求，您可以设置适当的审核规则，帮助您识别出含有敏感词、不当内容的音频，及时发现并处理不合规的内容。

  场景优势：
  - 审核效率高：减少人工审核的工作量，提高教学内容的准确性，避免出现错误或不当的言论。

- 准确率高：过滤掉不良信息和不当言论，保证教学内容安全。

# 4 使用限制

## 文本内容审核（V3）

- 支持“亚太-新加坡”区域。
- 待检测文本的编码格式为“utf-8”，限定1500个字符以内，文本长度超过1500个字符时，只检测前1500个字符。
- 默认API调用最大并发为50（表示1秒内最多请求50次），如需调整更高并发限制请通过**工单**联系专业工程师为您服务。
- 目前国际站策略相对于中国站较宽松，具体详情可通过工单联系华为云技术人员。

## 文本内容审核（V2）

- 支持“中国-香港、亚太-新加坡”区域。
- 待检测文本的编码格式为“utf-8”，限定5000个字符以内，文本长度超过5000个字符时，只检测前5000个字符。
- “中国-香港、亚太-新加坡”默认API调用最大并发为5（表示1秒内最多请求5次），如需调整更高并发限制请通过**工单**联系专业工程师为您服务。

## 图像内容审核（V3）

- 支持“亚太-新加坡”区域。
- 支持识别处理JPG、PNG、JPEG、WEBP、GIF、TIFF、TIF、HEIF格式的图片。
- 图像各边的像素大小在20到6000px之间。
- 图片base64编码后大小不超过10MB（原图像大小不超过7.5MB）。
- 默认API调用最大并发为10（表示1秒内最多请求10次），如需调整更高并发限制请通过**工单**联系专业工程师为您服务。
- 目前国际站策略相对于中国站较宽松，具体详情可通过工单联系华为云技术人员。

## 图像内容审核（V2）

- 支持“中国-香港、亚太-新加坡”区域。
- 支持识别处理PNG、JPEG、BMP、WEBP、GIF格式的图片。
- 图像各边的像素大小在10到10000px之间。

- 图片base64编码后大小不超过10MB（原图像大小不超过7.5MB）。
- “中国-香港、亚太-新加坡”默认API调用最大并发为1（表示1秒内最多请求1次），如需调整更高并发限制请通过**工单**联系专业工程师为您服务。

## 音频内容审核

- 支持“亚太-新加坡”区域。
- 支持WAV、MP3、AAC、AMR、3GP、M4A、WMA、OGG、APE、FLAC、ALAC、WAVPACK、SILK_V3格式的音频文件。
- 音频文件大小不超过200MB。
- 默认API调用最大并发为10（表示1秒内最多请求10次），如需调整更高并发限制请通过**工单**联系专业工程师为您服务。

## 视频内容审核

- 支持“亚太-新加坡”区域。
- 支持AVI、FLV、MP4、MPG、WMV、MOV、WMA、RMVB、m3u8等格式。
- 视频文件大小不超过300Mb，视频时长小于等于2小时。
- 默认API调用最大并发为10（表示1秒内最多请求10次），如需调整更高并发限制请通过**工单**联系专业工程师为您服务。
- 目前国际站策略相对于中国站较宽松，具体详情可通过工单联系华为云技术人员。

## 音频流内容审核

- 支持“亚太-新加坡”区域。
- 音频流url地址，支持rtmp、rtmps、hls、http、https等主流协议。
- 默认API调用最大并发为10（表示1秒内最多请求10次），如需调整更高并发限制请通过**工单**联系专业工程师为您服务。

# 5 与其他云服务的关系

## 统一身份认证服务

统一身份认证（Identity and Access Management，简称IAM）服务，IAM为内容审核提供了用户认证和鉴权功能。IAM的更多信息请参见《统一身份认证服务用户指南》。

## 云监控

云监控（Cloud Eye）可以监控内容审核的相关指标，用户可以通过指标及时了解内容审核各服务的使用情况。Cloud Eye更多信息请参见《云监控用户指南》。监控指标说明及查看操作请参见**监控指标**。

## 对象存储服务

对象存储服务（Object Storage Service，简称OBS）是稳定、安全、高效、易用的云存储服务。内容审核大多数接口都涉及到对用户的数据处理，用户的大量数据采用OBS批量方式处理，可以提升云上的处理的总体效率。

内容审核部分接口支持从OBS上采用临时授权或者匿名公开授权的方式获取数据并进行处理。OBS更多信息请参见《对象存储服务API参考》和《对象存储服务开发指南》。

# 6 使用服务

内容审核提供了Web化的服务管理平台，即管理控制台，以及基于HTTPS请求的API管理方式。

- 您可以在管理控制台申请开通内容审核服务、查看服务的调用成功和失败次数。
- 内容审核以开放API的方式提供给用户，用户可以将内容审核集成到第三方系统调用API。

具体流程如下：

**步骤1** 申请服务

用户可通过管理控制台申请服务，申请服务的具体操作步骤请参见“**申请服务**”章节。

**说明**

- 服务只需要开通一次即可，后面使用时无需再申请。
- 本服务暂时仅面向企业用户开放，个人用户暂不支持开通。

**步骤2** 获取请求认证

调用内容审核的API有如下两种认证方式，请任选其中一种进行认证鉴权。

- Token认证：通过Token认证调用请求，具体操作请参见**Token认证**。
- AK/SK认证：通过AK/SK加密调用请求。AK/SK认证安全性更高，具体操作请参见**AK/SK认证**。

**步骤3** 调用API

内容审核以API的方式提供服务，具体操作请参见**《内容审核API参考》**。

**步骤4** 查看服务使用信息

- 您可以在内容审核控制台查看服务调用总次数。
- 您可以通过单击页面中的“查看监控指标”，在控制台查看服务调用成功的次数和失败的次数等历史数据。

**----结束**

# 7 监控指标

## 功能说明

本节定义了内容审核上报云监控的监控指标的命名空间，监控指标列表和维度定义，用户可以通过查看监控指标和云监控提供的API接口来检索内容审核产生的监控指标。

## 命名空间

SYS.MODERATION

## 内容审核监控指标

**表 7-1** 内容审核支持的监控指标

| 指标ID | 指标名称 | 指标含义 | 取值范围 | 测量对象 | 监控周期（原始指标） |
| --- | --- | --- | --- | --- | --- |
| successful_call_times_of_service | 调用内容审核成功次数 | 该指标用于统计调用服务成功次数。<br>单位：次/分钟 | ≥ 0 times/min | 内容审核 | 1分钟 |
| failed_call_times_of_service | 调用内容审核失败次数 | 该指标用于统计调用服务失败次数。<br>单位：次/分钟 | ≥ 0 times/min | 内容审核 | 1分钟 |

**说明**

每个子服务都有调用成功次数和失败次数两个指标。

## 维度

**表 7-2** 维度说明

| Key | Value |
| --- | --- |
| call_of_interface | 接口 |

## 查看监控指标

内容审核控制台只记录服务调用总次数，您可以通过公有云平台提供的云监控管理控制台，直观地查看服务调用成功和失败的次数。

1. 登录管理控制台。
2. 选择“人工智能 > 内容审核”，进入“内容审核”界面。
3. 在左侧导航栏单击已开通并调用的服务，进入对应服务详情页面。
4. 单击“查看监控指标”，进入云监控控制台查看服务调用成功和失败的次数等具体信息。

   您可以通过选择时长，查看对应时间的监控数据。当前支持查看“近1小时”、“近3小时”、“近12小时”、“24小时”、“近七天”的监控数据。

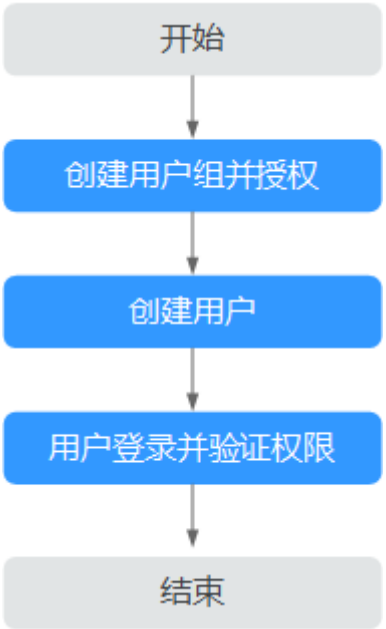
  

# 8 （可选）授权子账号使用内容审核服务

本章节通过简单的用户组授权方法，将内容审核对应区域的“Tenant Guest”权限和对象存储的“OBS Buckets Viewer”策略授予用户组，并将用户添加至用户组中，从而使子账户拥有对应的操作权限，操作流程如图8-1所示。

## 示例流程

**图 8-1** 给用户授权内容审核权限流程




## 步骤 1：创建用户组并授权

用户组是用户的集合，IAM通过用户组功能实现用户的授权。您在IAM中创建的用户，需要加入特定用户组后，用户才具备用户组所拥有的权限。关于创建用户组并给用户组授权的方法，可以参考如下操作。

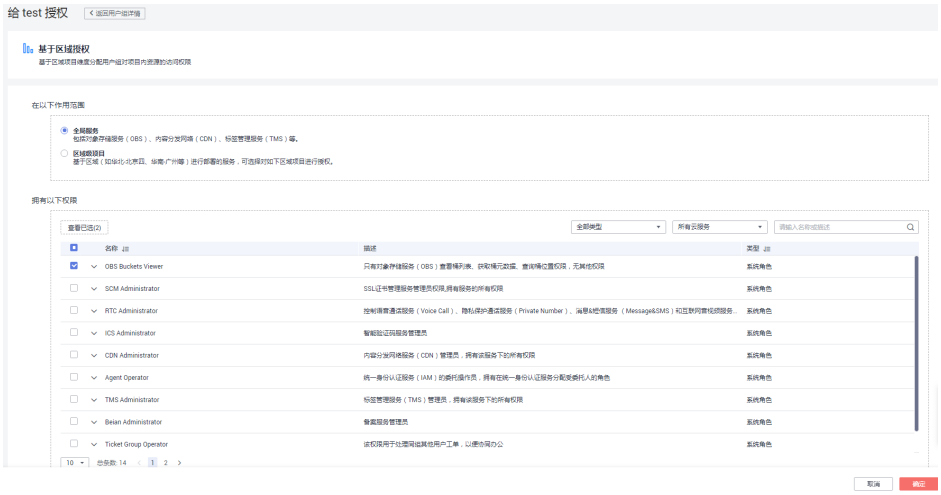
1. 使用注册的华为账号登录华为云，登录时请选择“账号登录”。
2. 进入华为云控制台， 鼠标移动至控制台页面中单击右上角的用户名，选择“统一身份认证”。

3. 在统一身份认证服务的左侧导航空格中，单击“用户组>创建用户组”。
4. 在“创建用户组”界面，输入“用户组名称”单击“确定”。

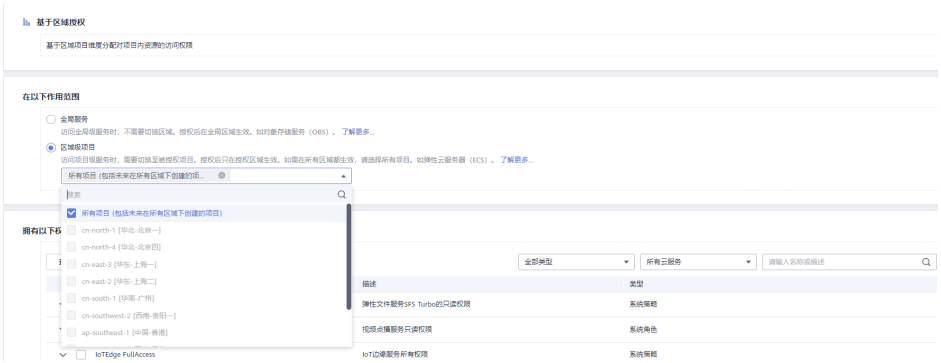
   用户组创建完成之后，界面自动返回用户组列表，列表中显示新建的用户组。
5. 进行全局服务配置，单击新建用户组右侧的“权限配置”，在“权限管理”页签中，单击列表左上方的“授权”。此处作用范围选择“全局服务”。勾选需要授予用户组的权限“Tenant Guest”和“OBS Buckets Viewer”，单击“确定”。如图8-2所示。

**图 8-2** 全局服务配置




6. 进行区域级权限配置，单击新建用户组右侧的“权限配置”，在“权限管理”页签中，单击列表左上方的“授权”，此处作用范围选择“区域级项目”，勾选所有项目 (包括未来在所有区域下创建的项目)。勾选需要授予用户组的权限“Tenant Guest”，单击“确定”，完成用户组授权。如图8-3所示。

**图 8-3** 区域级权限配置




7. 返回用户组列表，单击新建用户组右侧的“权限配置”，在“权限管理”页签中查看已经配置好的权限。如图8-4所示。

**图 8-4** 权限管理


## 步骤 2：创建 IAM 用户

IAM用户与企业中的实际员工或是应用程序相对应，有唯一的安全凭证，可以通过加入一个或多个用户来获得用户组的权限。关于IAM用户的创建方式请参见如下步骤。

1. 在统一身份认证服务，左侧导航中，单击“用户>创建用户”。
2. 在“创建用户”界面中填写参数信息，完成后单击“下一步”。具体参数说明请参见**创建IAM用户**。
3. 在界面中填写参数信息，单击“确定”，完成用户创建。
4. 为**用户组添加用户**，使用户具备用户组的权限，实现用户的授权。

## 步骤 3：用户登录并验证权限

用户创建完成后，可以使用新用户的用户名及身份凭证登录华为云验证权限。

1. 在华为云登录页面，单击右下角的“IAM用户登录”.
2. 在“IAM用户登录”页面，输入账号名、用户名及用户密码，使用新创建的用户登录。
   - 账号名为该IAM用户所属华为云账号的名称。
   - 用户名和密码为账号在IAM创建用户时输入的用户名和密码。
   - 如果登录失败，您可以联系您的账号主体，确认用户名及密码是否正确，或是重置用户名及密码。
3. 登录成功后，进入华为云控制台，登录后默认区域为“华为-北京四”，请先切换至授权区域。
4. 在“服务列表”中选择内容审核，在服务管理页面进行OBS授权、开通服务、调用均能正常使用，则表示授权已生效。

# 9 计费说明

## 计费项

目前商用的服务计费方式有两种，按需付费和折扣套餐包两种方式计费。了解内容审核价格详情，请参见**内容审核价格详情**。

## 计费模式

内容审核已商用的服务提供两种计费模式供您选择：按需计费和套餐包计费。

- **按需计费**

按需计费按照调用次数阶梯价格计费，按月累计，一个自然月后次数清零重新累计。促销活动期间针对不同服务，每个用户每月有对应的免费调用次数，具体计费价格详情请参见**内容审核价格详情**。

**说明**

- 只有调用成功才会计算调用次数，未用完的免费调用次数不流转到下一个月。
- 计费规则：调用次数（审核的图片数量）阶梯计费，每审核一张图片记为一次调用，按月累计，一个自然月后调用次数清零重新累计。
- 计费周期：按小时计费，实时扣费（账单出账时间通常在当前计费周期结束后一小时内，具体出账时间以系统为准）。

- **套餐包计费**

您可以购买套餐包，扣费时调用次数会先在套餐包内进行抵扣，抵扣完后的剩余调用量默认转回按需计费方式。具体计费价格详情请参见**内容审核价格详情**。这种购买方式相对于按需付费提供更大的折扣，对于长期使用者，推荐该方式。

购买前需要了解以下几点：

1. 确定购买时长和购买数量后，系统会自动计算出配置费用。
2. 套餐包支持多个购买，可叠加使用。

   例如：如果调用次数不够用了，每月60万次要想升级到每月120万次，只需再购买一个60万次的套餐包即可，支持叠加使用。可购买的套餐包规格，请参考**内容审核价格详情**。
3. 套餐包费用为一次性支付，即刻生效，暂不支持指定日期生效，需在套餐包生效期内使用，到期自动结束。

4. 套餐包的剩余次数无法叠加到下个周期内，请您在套餐包到期前使用。

   例如：您1月1日购买了1个周期一年的套餐包，则该套餐在次年1月1日会自动结束，即使您在该有效期内未调用内容审核服务，该套餐也不会延期或叠加到下个周期内，且无法退还费用。

5. 超过套餐包内额度的部分转按需计费后，按当月累计调用量落入的阶梯计费。

**说明**

- 套餐包余量可登录内容审核控制台在“费用与成本 > 我的套餐”中查看。


## 欠费

按需购买的接口是按照每小时扣费，当账户的余额不足时，无法对上一个小时的费用进行扣费，就会导致欠费。

您续费后可继续正常使用，请注意在保留期进行的续费，是以原到期时间作为生效时间，您应当支付从进入保留期开始到续费时的服务费用。

**说明**

您账号欠费后，会导致部分操作受限，建议您尽快续费。具体受限操作如下：

- 按需计费方式购买的API接口不可调用。
- 套餐包方式购买的API接口，在欠费后如果套餐包内有剩余，可继续使用，但不可以再次购买和续期。
- 无法开通服务。

## 续费

资源包到期后，您可以进行续费以延长资源包的有效期也可以设置到期自动续费。

## 服务到期

- 包年包月资源包到期后，自动转为按需计费。
- 保留期满仍未续订或充值，数据将被删除且无法恢复。