ALIBABA CLOUD

阿里云

内容安全

用户指南

文档版本：20220524

阿里云

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。

2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。

3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。

4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。

5. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含“阿里云”、“Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。

6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

格式	说明	样例
危险	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	危险 重置操作将丢失用户配置数据。
警告	该类警示信息可能会导致系统重大变更甚至故障，或者导致人身伤害等结果。	警告 重启操作将导致业务中断，恢复业务时间约十分钟。
注意	用于警示信息、补充说明等，是用户必须了解的内容。	注意 权重设置为0，该服务器不会再接受新请求。
说明	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	说明 您也可以通过按Ctrl+A选中全部文件。
>	多级菜单递进。	单击**设置**> **网络**> **设置网络类型**。
粗体	表示按键、菜单、页面名称等UI元素。	在**结果确认**页面，单击**确定**。
Courier字体	命令或代码。	执行 `cd /d C:/window` 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	`bae log list --instanceid` *Instance_ID*
[] 或者 [a|b]	表示可选项，至多选择一个。	`ipconfig [-all|-t]`
{} 或者 {a|b}	表示必选项，至多选择一个。	`switch {active|stand}`

目录

1.OSS违规检测

1.1. 使用简介

OSS违规检测能够检测阿里云对象存储OSS服务中的图片、视频、语音是否包含色情、涉政等违规内容，并可以自动冻结检测出的违规内容（禁止通过公网访问这些违规内容），帮助您规避可能遇到的内容违规风险。

功能介绍

OSS违规检测支持检测指定OSS Bucket中的增量内容和存量内容，即增量扫描和存量扫描。该功能无需您开发，只需要少许页面配置即可接入使用。

- 增量扫描

  开启增量扫描后，当您在OSS Bucket上传了新的图片、视频、语音时，内容安全将自动检测新增的图片、视频是否存在违规。增量扫描一次配置即可长期生效。

- 存量扫描

  存量扫描目前不支持自动检测，需要您手动创建扫描任务。创建后，内容安全将在指定时间，一次性扫描您指定的OSS Bucket已有的图片、视频、语音文件是否存在违规。

使用限制

OSS违规检测只向开通了阿里云对象存储OSS（Object Storage Service）的用户提供服务。关于OSS服务的更多信息，请参见什么是对象存储OSS。

使用OSS违规检测前，您必须在内容安全控制台完成使用授权，授予内容安全服务对您的OSS Bucket的访问权限。

> **说明** 您在首次登录内容安全控制台并访问OSS违规检测页面时，可以一键完成该操作。具体操作，请参见授权内容安全访问OSS存储空间。

OSS违规检测说明

对象	说明
OSS Bucket的地域	OSS违规检测目前支持扫描以下地域的OSS Bucket： • `oss-cn-hangzhou.aliyuncs.com`：表示华东1（杭州）地域的OSS Bcuket。 • `oss-cn-shanghai.aliyuncs.com`：表示华东2（上海）地域的OSS Bcuket。 • `oss-cn-qingdao.aliyuncs.com`：表示华北1（青岛）地域的OSS Bcuket。 • `oss-cn-beijing.aliyuncs.com`：表示华北2（北京）地域的OSS Bcuket。 • `oss-cn-shenzhen.aliyuncs.com`：表示华南1（深圳）地域的OSS Bcuket。 **说明** 关于地域的具体介绍，请参见访问域名和数据中心。如果您的OSS Bucket位于其他地域，则暂时无法使用OSS违规检测。
检测场景	OSS违规检测支持的检测场景如下： • 图片违规检测场景：涉黄、涉政暴恐、图文违规、不良场景。 • 视频违规检测场景：涉黄、涉政暴恐、图文违规、不良场景、视频语音违规。 • 语音违规检测场景：涉黄、涉政暴恐、广告、垃圾信息、辱骂、灌水、违禁、无意义、呻吟声。

对象	说明
检测对象	OSS违规检测支持检测的图片、视频文件格式如下： • 图片文件格式：*JPG*、*PNG*、*JPEG*、*GIF*、*BMP*、*WEBP*。 • 视频文件格式：*AVI*、*MP4*、*3GP*、*MKV*、*MPG*、*MPEG*、*TS*、*RMVB*、*WMV*、*FLV*、*MOV*。 • 语音支持格式：*MP3*、*WAV*、*AAC*、*WMA*、*OGG*、*M4A*、*AMR*、*AUDIO*。

使用流程

步骤	操作	详细说明
步骤一	授权内容安全访问OSS存储空间	OSS违规检测只向开通了阿里云对象存储OSS服务的用户提供服务，所以请确保您已开通OSS存储空间并授权内容安全读取OSS Bucket。 关于授权的具体操作，请参见授权内容安全访问OSS存储空间。 **说明** 首次使用OSS违规检测服务时，需要先授权OSS访问权限。服务只需要授权一次即可，完成授权后您无需再次授权。
步骤二	配置回调通知	内容安全支持以异步消息通知的方式向您发送机器内容识别和您自助审核的结果。若您的业务需要使用或集成相关数据，您需要配置回调通知。 关于回调通知的具体操作，请参见配置消息通知。
步骤三	配置风险库	为了使检测结果更贴合您的实际业务，内容安全也支持自定义图库。您可以使用自定义图库指定需要拦截、放行、人工审核的图片。 关于风险库的具体操作，请参见配置风险库。
步骤四	增量扫描	通过增量扫描设置，您可以让内容安全对指定OSS Bucket中新增的图片、视频自动进行违规检测（每当Bucket中有新增内容，将自动触发扫描），并实时查看近7天的增量扫描结果和数据统计。 具体操作如下： 1. 设置增量扫描任务 2. 查看统计数据 3. 查看扫描结果
	存量扫描	对于手动创建一个存量扫描任务，指定OSS Bucket中的已有图片或视频文件进行一次性违规检测，并在检测完成后查看扫描结果和数据统计信息。 具体操作如下： 1. 创建存量扫描任务 2. 查看统计数据 3. 查看扫描结果

1.2. 授权内容安全访问OSS存储空间

OSS违规检测只向开通了阿里云对象存储OSS服务的用户提供服务。在使用OSS违规检测前，您需要为内容安全授予访问OSS存储空间的权限，才能对存储在OSS空间中的文件进行检测。本文介绍了如何授权内容安全访问OSS存储空间。

> **说明** 防盗链功能通过设置Referer白名单以及是否允许空Referer，限制仅白名单中的域名可以访问您存储空间内的资源。如果您已开启OSS防盗链，则必须在OSS控制台将 `https://yundun.console.aliyun.com` 增加至Referer白名单。具体操作，请参见防盗链。

前提条件

已开通对象存储OSS服务。更多信息，请参见开通对象存储OSS服务。

操作步骤

1. 登录内容安全控制台。
2. 在左侧导航栏，选择**设置 > OSS违规检测**或者**OSS违规检测 > 存量扫描**。
3. 根据页面提示，单击**去授权**。
4. 在**云资源访问授权**页面，单击**同意授权**，完成授权。

后续步骤

完成授权后，您就可以使用OSS违规检测功能，在使用OSS违规检测功能之前，您需要先配置消息通知。更多信息，请参见配置消息通知。

如果您需要使检测结果更贴合您的实际业务，您可以配置风险库，指定您需要拦截、放行、人工审核的图片。更多信息，请参见配置风险库。

1.3. 配置消息通知

内容安全支持以异步消息通知的方式向您发送机器内容识别和您自助审核的结果。若您的业务需要使用或集成相关数据，您需要配置消息通知。本文介绍了配置OSS违规检测消息通知的具体操作。

相关概念

在配置回调通知前，请了解下表中描述的相关概念。

名称	说明
回调地址	回调地址是您在内容安全控制台配置的服务端地址，通常是您自己的业务服务器的公网地址。回调地址需要满足以下要求： • 应为HTTP、HTTPS协议接口的公网可访问的URL。 • 支持POST方法。 • 支持传输数据编码采用UTF-8。 • 支持数据接收格式为 `application` 、 `x-www-form-urlencoded` 。 • 支持表单参数checksum和content。
Seed	Seed值用于校验发送到您设置的HTTP回调通知地址的请求是否来自内容安全服务端。

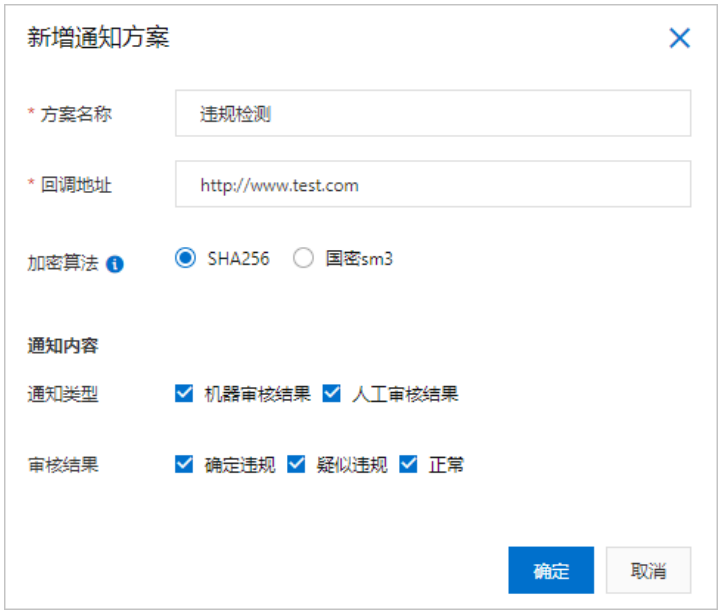
名称	说明
回调次数	您的服务端接收到内容安全推送的回调消息后，如果返回的HTTP状态码为200，表示接收成功；如果返回其他的HTTP状态码，均视为接收失败。如果接收失败，内容安全服务端会重复推送回调消息，直至您接收成功。内容安全最多重复推送回调消息16次。
回调数据	回调数据是内容安全服务端向您设置的回调通知地址返回的数据内容。关于回调数据的结构描述，请参见回调通知表单数据。

回调通知表单数据

名称	类型	描述
checksum	String	总和校验码，由 `<用户uid> + <seed> + <content>` 拼成字符串，通过SHA256算法生成。用户UID即阿里云账号ID，可以在阿里云控制台上查询。 说明 为防篡改，您可以在获取到推送结果时，按上述算法生成字符串，与checksum做一次校验。
content	String	字符串格式保存的JSON对象，请自行解析反转成JSON对象。关于content解析成JSON后的结构，请参见。 说明 在内容检测API和OSS违规检测中，content的参数结构不同。

设置消息通知

1. 登录内容安全控制台。
2. 在左侧导航栏，选择**设置 > OSS违规检测**。
3. 在**OSS违规检测**页面，单击**消息通知**页签。
4. 在**消息通知**页签下，单击**新增通知方案**。



5. 在**新增通知方案**页面，完成回调通知配置，并单击**确定**。

   回调通知配置的描述如下。

配置项	说明
方案名称	设置方案的名称。必须在12个字符以内，包含中英文、下划线和数字。
回调地址	填入回调通知地址。
加密算法	设置对回调通知内容进行加密的算法。内容安全会将返回结果（由`用户uid + seed + content`拼接的字符串）按照您设置的加密算法加密后，再发送到您的回调通知地址。加密算法分为以下两类： ○ **SHA256**（默认）：使用HMAC-SHA256加密算法。 ○ **国密SM3**：使用国密SM3加密算法，返回十六进制的字符串，且字符串由小写字母和数字组成。 例如，回调通知**abc**经国密SM3加密后返回**66c7f0f462eeedd9d1f2d46bdc10e4e24167c4875cf2f7a2297da02b8f4ba8e0**。
通知类型	开启扫描回调后，选择对哪些类型的扫描结果进行回调通知。支持多选，通知类型分为以下两种： ○ **机器审核结果** ○ **自助审核结果**
审核结果	开启扫描回调后，选择对哪些类型的审核结果进行回调通知。支持多选，审核结果分为以下三种： ○ **确定违规** ○ **疑似违规** ○ **正常**

设置完成后，系统自动生成Seed。Seed值用于校验您的回调接口收到的请求来自阿里云。请保存自动生成的Seed，根据需要应用于您的业务。

content表单字段结构说明

启用回调通知后，内容安全将按照回调配置发送OSS违规检测的回调通知。回调通知中包含content表单数据。下表描述了content表单字段的结构。

名称	类型	是否必选	说明
bucket	String	是	OSS Bucket的名称。
object	String	是	OSS文件名。
stock	Boolean	是	是否是存量内容，取值： • *true*：表示是存量内容。 • *false*：表示是增量内容。
region	String	是	OSS文件所在地域。
freezed	Boolean	是	内容是否被冻结（禁止通过公网访问这些内容），取值： • *true*：表示被冻结。 • *false*：表示未被冻结。

名称	类型	是否必选	说明
scanResult	JSONObject	否	扫描结果。根据检测对象（图片、视频）的不同，结构有差异。 • 针对图片对象，结构与图片检测中的results的返回参数相同。具体信息，请参见图片同步检测。 • 针对视频对象，结构与视频异步检测中的results的返回参数相同。具体信息，请参见视频异步检测。
auditResult	JSONObject	否	自助审核结果，执行自助审核操作时才会有该字段。具体结构描述，请参见auditResult。 **说明** 当您选择通知类型为**自助审核结果**，才出现该字段。

auditResult

名称	类型	是否必选	说明
suggestion	String	是	自助审核的结果。取值： • *block*：审核时设置违规 • *pass*：审核时设置正常
resoureStatus	Integer	是	自助审核后，object的状态。取值： • *0*：已删除 • *1*：已冻结 • *2*：可用可访问

content示例

```
{
    "bucket": "xxxxb",
    "freezed": false,
    "object": "xxx.jpg",
    "region": "cn-hangzhou",
    "scanResult": {
        "code": 200,
        "dataId": "5C919E2FBD6CD6940D4A8B46",
        "msg": "OK",
        "results": [
            {
                "label": "porn",
                "rate": 100.0,
                "scene": "porn",
                "suggestion": "block"
            },
            {
                "label": "normal",
                "rate": 99.6,
                "scene": "terrorism",
                "suggestion": "pass"
            }
        ],
        "taskId": "img1ZxzVQUuAz36scZEjyGXzx-1qpzhY"
    },
    "auditResult": {
        "resoureStatus": 2,
        "suggestion": "pass"
    },
    "stock": false
}
```


后续步骤

完成配置消息通知后，您就可以使用OSS违规检测的存量扫描和增量扫描功能。详细信息，请参见设置增量扫描任务、创建存量扫描任务。

如果希望检测结果更贴合实际业务，您可以使用自定义图库针对性地拦截、放行、自助审核的图片，应对突发的管控需求。详细信息，请参见配置风险库。

1.4. 配置风险库

为了使检测结果更贴合您的实际业务，内容安全支持自定义图库。自定义图库适用于图片和视频鉴黄、图片和视频涉政暴恐识别场景。您也可以使用自定义图库对指定的图片进行拦截、放行和人工审核，以便于应对突发的管控需求。

背景信息

根据用途不同，自定义图库分为黑名单、白名单、疑似名单。在检测中应用自定义图库后，若被检测图片命中图库中的样本，则会被打上图库对应的识别结果标签。黑名单图库对应的识别结果是违规（拦截），白名单对应正常（放行），疑似名单则对应疑似（人工审核）。自定义图库包括系统回流图库和用户创建图库。

- 系统回流图库由您的自助审核记录自动生成，默认应用于所有同类场景的图片检测。您可以管理系统回流图库中的图片，不可以操作系统回流图库，例如停用或删除图库。
- 用户创建图库由您自行添加，可用于某次检测或某类检测场景。您可以管理用户创建图库中的图片，也可以操作用户创建图库。

> **说明** 您可以创建10个自定义图库（不含系统回流图库），且在每个图库中添加最多10000张样本图片。

操作步骤

1. 登录内容安全控制台。
2. 在左侧导航栏，选择**OSS违规检测 > 风险库管理**。
3. 在**风险库管理**页面，单击**创建图库**。
4. 在**创建图库**对话框中，请参考下表完成图库配置，并单击**确定**。

   创建图库参数说明表

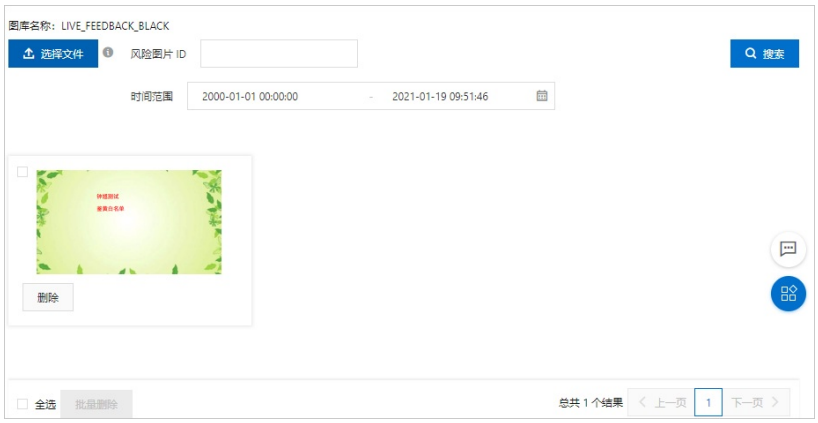
配置项	说明
名称	为图库命名。名称不超过64个字符。建议设置为可读性较强的中文名称。
使用场景	选择图库的使用场景，取值： ○ **鉴黄**：图片、视频鉴黄（scenes包含porn）。 ○ **暴恐**：图片、视频暴恐涉政识别（scenes包含terrorism）。 ○ **广告**：图片、视频图文违规识别（scenes包含ad）。 ○ **不良场景**：图片、视频不良场景识别（scenes包含live）。
识别结果	选择图库的用途，取值： ○ **黑名单**：若命中图库中样本，则机审结果返回违规。 ○ **疑似名单**：若命中图库中样本，则机审结果返回疑似。 ○ **白名单**：若命中图库中样本，则机审结果返回正常。
BizType	BizType允许您根据不同的业务需求配置并应用不同的图库，例如使用BizType指定在某次检测中应用图库。BizType生效逻辑如下： ○ 图库设置BizType为“A”，检测时API中传递了BizType为“A”，则检测时只会使用BizType为“A”的图库（前提是图库已启用）。 ○ 其他情况下，检测均会使用所有已启用的图库。 **说明** 包含Biztype属性的风险图库不适用于OSS违规检测。

   成功创建图库，已创建图库默认启用。

5. 在**自定义图库**页签，单击指定图库右侧操作列下的**管理**。

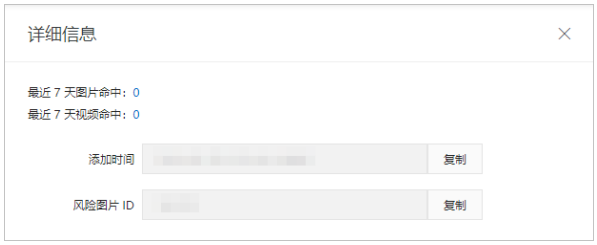
   自定义图库页签下显示所有自定义图库，其中系统回流图库的名称前有**系统**标识，且按照使用场景_FEEDBACK_WHITE/BLACK命名。例如，**LIVE_FEEDBACK_BLACK**是由系统回流生成的用于不良场景的黑名单图库。

6. 在**图库管理**页面，您可以根据需要执行如下操作。



> **说明** 在图库中新增、删除图片样本，大约需要15分钟生效。

- 根据**风险图片ID**、**时间范围**查询图片。
- 单击图片，展开图片的**详细信息**面板，查看近期命中数量、添加时间、风险图片ID等信息。



- 单击**选择文件**，上传图片到图库。

> **说明** 支持上传PNG、JPG、JEPG和BMP格式的图片。

- 单击图片下的**删除**，删除图片；选中多张图片后，单击**批量删除**，批量删除图片。

相关操作

对于自定义图库（非系统回流图库），您可通过**自定义图库**页签中**操作**列下的**删除**、**编辑**和**停用**对目标图库进行操作。

后续步骤

完成配置风险库后，您就可以使用OSS违规检测功能，在使用OSS违规检测功能之前，您需要先配置消息通知。请参见配置消息通知。

1.5. 增量扫描

1.5.1. 设置增量扫描任务

通过增量扫描设置，您可以对指定OSS Bucket中新增的图片、视频自动进行违规检测（每当Bucket中有新增内容，将自动触发扫描），并实时查看最近7天的增量扫描结果。

前提条件

- 已经开通OSS对象存储服务并授权内容安全读取您的OSS Bucket。具体操作，请参见授权内容安全访问OSS存储空间。
- 已完成配置回调通知。具体操作，请参见配置消息通知。

背景信息

关于OSS违规检测的使用限制（例如支持的OSS Bucket、检测场景、文件格式等），请参见使用限制。

操作步骤

1. 登录内容安全控制台。
2. 在左侧导航栏，选择**设置 > OSS违规检测**。
3. 在**OSS违规检测**页面的**增量扫描**页签，按照向导完成如下配置：

i. **选择Bucket**。

从左侧**待选择**框，选中需要检测的Bucket，添加到右侧的**已选择**框。

待选择框罗列了当前阿里云账号在OSS空间中可检测到的所有Bucket。

然后单击**下一步**。

ii. **设置过滤条件**。

配置说明：

过滤或排除：默认扫描已选Bucket中的所有文件，如果您需要指定扫描Bucket中的哪些文件或者不要扫描哪些文件，则可以单击Bucket右侧的**设置过滤条件**，并在**过滤条件**对话框，按照页面提示完成过滤条件配置，然后单击**确定**。

过滤条件参数说明

参数	说明
过滤条件	支持**包含**（表示扫描指定的路径）或**排除**（表示扫描时排除指定的路径）。 必须且只能选择一种方式。
路径	输入要包含或排除的路径。 单击**添加**可以添加多个路径。最多允许添加10个。 配置示例（以**包含**条件为例）： ■ 添加 `img/test_`，表示只扫描Bucket中以*img/test_*为前缀的文件。 ■ 如果要扫描的文件在特定目录下，您可以在文件名前加上目录路径，以整体作为前缀。例如，您要扫描的文件在*img/202011*目录下，且前缀为*test_*，您可以添加 `img/202011/test_` 作为路径。

然后单击**下一步**。

iii. **配置检测场景。**

OSS违规检测为您提供8个推荐配置，每种推荐配置包含大类场景（图片、视频、语音）和细分场景（每种推荐配置的细分场景均不相同）。如果推荐的配置与您业务不符，您可以关闭或者开启大类场景，但不支持修改细分场景。

然后单击**下一步**。

iv. **配置扫描范围**

扫描范围根据您配置的检测场景显示。例如，您在**场景配置**页面只勾选**图片**，那么**扫描配置**页面就只显示图片的配置信息。如果当前配置与您业务不符，您可以自定义扫描范围。

> **说明** 至少选择一类检测场景进行扫描，即图片、视频、语音中至少开启一项。

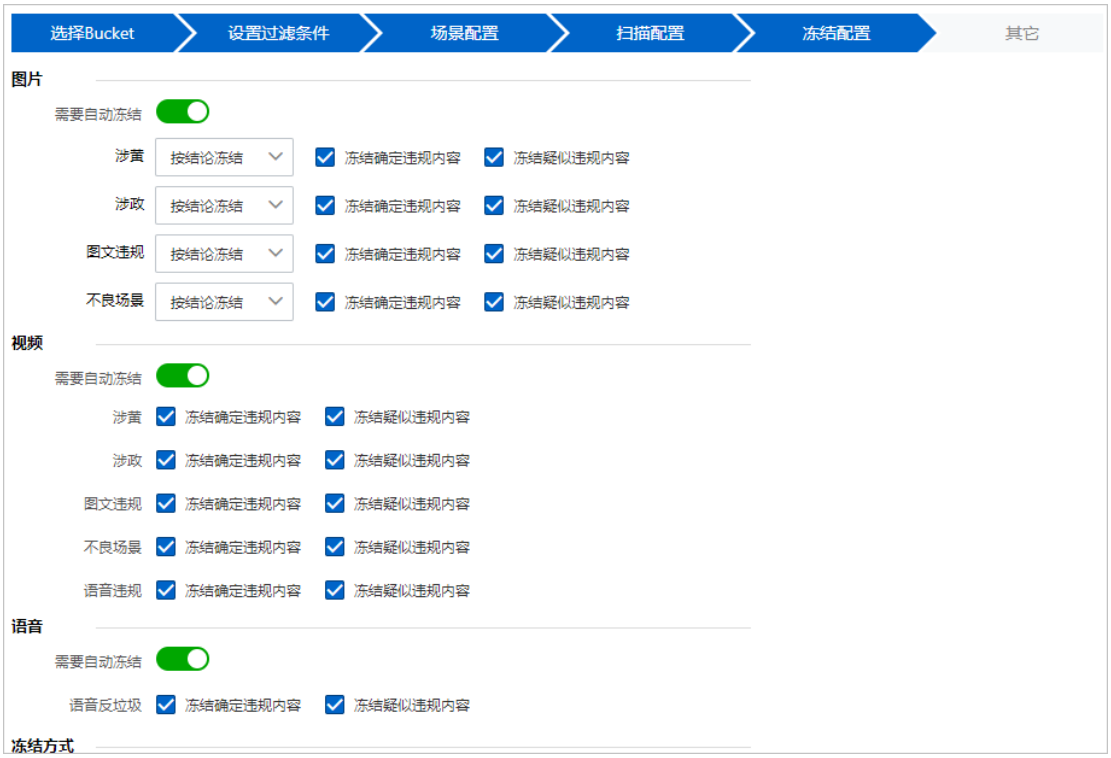
扫描配置参数说明

扫描范围	参数名称	说明
图片	**每日图片扫描上限**	每日扫描的图片张数默认为10000张。如果设置了扫描上限，扫描数量超出限制后将会停止扫描，因此会存在巨大违规图片外露的风险。常规情况下，不建议您设置扫描上限。
	检测无后缀文件	开启**检测无后缀文件**后，会扫描无后缀的图片文件，检测速度会降低，根据HTTP头的content-type判断是否为图片，支持content-type有：image/jpg，image/jpeg，image/png，image/gif，mage/bmp，image/webp。
视频	**每日视频扫描上限**	每日扫描的视频个数默认为1000个。如果设置了扫描上限，扫描数量超出限制后将会停止扫描，因此会存在巨大违规视频外露的风险。常规情况下，不建议您设置扫描上限。
	截帧频率	截帧频率默认为1帧/秒，您可以设置截帧频率（每多少秒截取一帧）。取值范围1~60帧/秒。截帧频率越高，识别准确率越高。
	单视频帧数上限	单视频帧数默认为200帧，您可以设置单个视频的最大截帧数量。取值范围：5~20000帧。
	单视频大小上限	单视频大小默认为500 MB，您可以设置单视频大小上限，取值范围：1~2048 MB，超过部分不会被检测。
语音	**每日语音扫描上限**	每日扫描的音频个数默认为1000个。如果设置了扫描上限，扫描数量超出限制后将会停止扫描，因此会存在巨大违规音频外露的风险。常规情况下，不建议您设置扫描上限。
	单音频大小上限	单音频大小默认为200 MB。您可以设置单音频大小上限，取值范围：1~2048 MB，超过部分不会被检测。

然后单击**下一步**。

v. **冻结配置**。

根据是否需要自动冻结检测出来的违规图片、视频和语音，配置对应的冻结规则，并选择一种**冻结方式**。



配置说明：

- 自动冻结违规图片：开启**图片**区域下的**需要自动冻结**开关，并为不同场景设置对应的冻结规则（不同场景的设置方法一样）。

  支持以下两种类型的冻结规则：

  - 按分值冻结：选择**按分值冻结**并设置一个违规分数阈值。当图片的违规分超过阈值时，将被冻结。

    > **注意** 请慎重修改冻结阈值。默认冻结分数是99.01分，正常情况下不建议设置到99分以下，分数过低可能会导致正常图片被冻结。

  - 按结论冻结：选择**按结论冻结**并选中要冻结的违规内容类型，可选项：**冻结确定违规内容**、**冻结疑似违规内容**。

    > **说明** 建议您优先使用按结论冻结，按分值冻结需具有一定的专业知识。如果您分值配置不适当可能会导致正常图片被冻结。

- 自动冻结违规视频：开启**视频**区域下的**需要自动冻结**开关，并为不同场景选择要冻结的违规内容类型，可选项：**冻结确定违规内容**、**冻结疑似违规内容**。
- 自动冻结违规语音：开启**语音**区域下的**需要自动冻结**开关，并为不同场景选择要冻结的违规内容类型，可选项：**冻结确定违规内容**、**冻结疑似违规内容**。

- 选择冻结方式：
  - **修改权限**：将您Bucket中public权限的违规文件设置为private访问权限。

    互联网用户无法读取private权限的文件，但您可以通过文件URL将私有文件分享给您的合作伙伴访问。更多信息，请参见在URL中包含签名。
  - **移动文件**：将您Bucket中违规的文件移动到Bucket中的备份目录（位置：*${bucket}/aligreen_freeze_backup/*），并删除原路径下的文件。

    > **注意** 选择该方式时请慎重。文件删除后将无法恢复，如需找回被删除的文件，您必须到备份目录中查找。

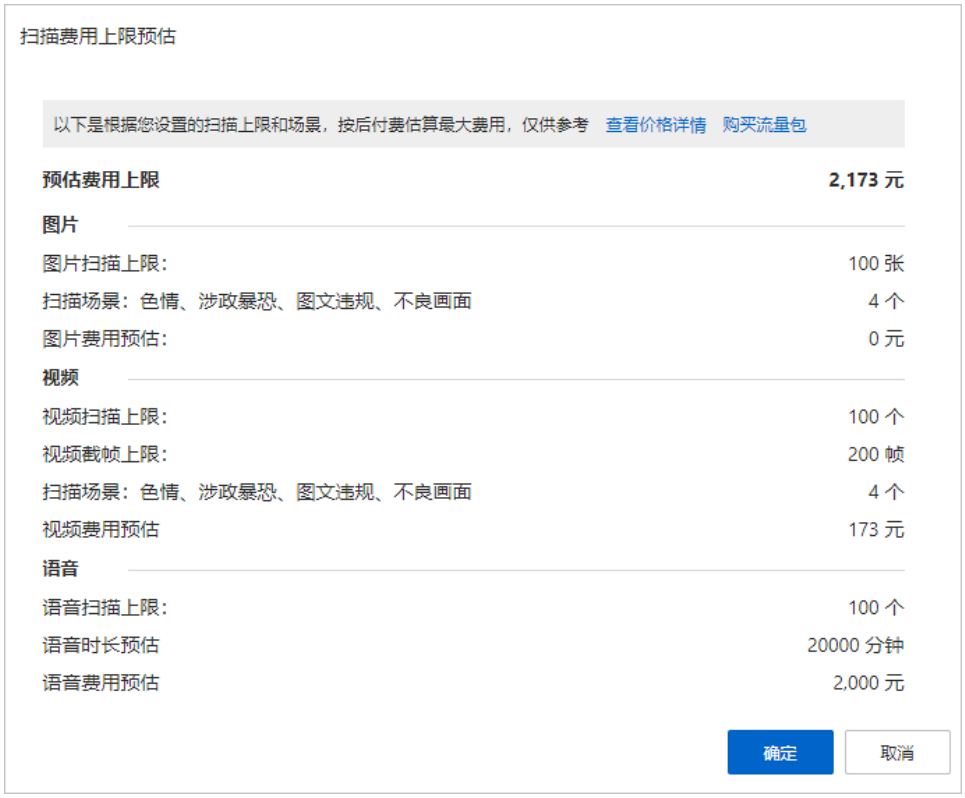
  然后单击**下一步**。

vi. （可选）**其它**。在**回调通知**区域，您可以选择一个已有的**回调通知方案**，通过指定的通知方案接收存量扫描的结果。

您必须先创建回调通知方案才可以进行设置。如果您未创建过回调通知方案，可以单击**新增通知方案**去添加通知方案。具体操作，请参见配置消息通知。

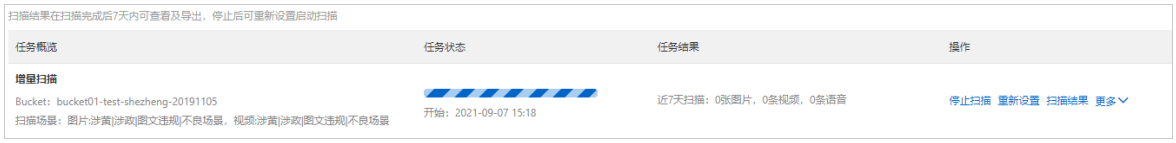
4. 在页面最下方，选中**我已经同意OSS违规检测服务条款**，并单击**保存**。

   OSS违规检测功能会根据您的配置为您预估出扫描费用的上限，您可以根据实际业务选择按量付费或者资源包抵扣方式。



然后单击**确定**。增量任务设置成功后，系统会自动跳转到**OSS违规检测 > 增量扫描任务**页面。

增量扫描配置保存后即时生效。系统会按照增量扫描配置，自动对已选择Bucket中新增的图片、视频进行违规检测。您可以在增量扫描任务列表中查看当前任务的状态。



OSS违规检测为您提供如下功能。您需要根据实际需要，选择合适的操作：

- **停止扫描**

  如果您不再需要扫描当前任务时，单击操作列**停止扫描**，停止扫描后新增的文件将不再执行扫描动作。

- **重新设置**

  如果您需要变更当前扫描任务时，单击操作列**重新设置**，在**设置 > OSS违规检测**页面重新设置增量扫描任务。

- **扫描结果**

  如果您需要查看当前任务的扫描结果，单击操作列**扫描结果**。您可以设置扫描的时间、任务等条件自定义查询结果信息。

  更多内容，请参见查看扫描结果。

- **数据统计**

  如果您需要查看当前任务的调用量统计信息，在操作列选择**更多 > 数据统计**，进入**OSS违规检测调用量**页面，查看最近7天的调用量。您可以设置扫描的时间、任务等条件自定义查询统计信息。

  更多内容，请参见查看统计数据。

- **查看配置**

  如果您需要查看当前任务的配置信息，在操作列选择**更多 > 查看配置**，展开**任务配置详情**面板查看。

- **用量说明**

如果您需要查看当前任务的用量，在操作列选择**更多 > 用量说明**，展开**任务用量说明**面板查看。

后续步骤

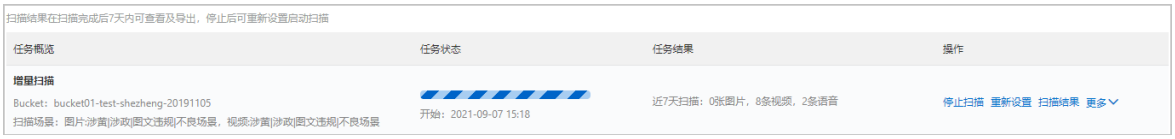
您可以在**OSS违规检测 > 增量扫描**页面查看近7天的扫描结果和数据统计信息。更多信息，请参见查看统计数据和查看扫描结果。

1.5.2. 查看扫描结果

OSS违规检测服务为您提供查看扫描结果的功能，当您完成增量扫描任务后，您可以随时在内容安全控制台查看扫描结果，并根据扫描结果执行自助审核。

操作步骤

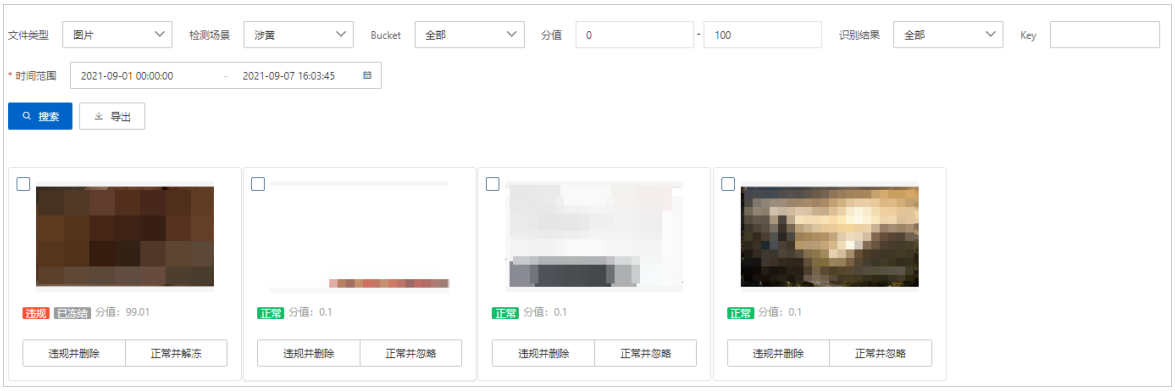
1. 登录内容安全控制台。
2. 在左侧导航栏，选择**OSS违规检测 > 增量扫描**。
3. 在**增量扫描**页面，查看您的**任务概览**、**任务状态**以及**任务结果**。



4. 单击操作列**扫描结果**，查询扫描结果并进行自助审核。

   默认显示最近7天的扫描结果和处理的违规内容。您可以设置**文件类型**、**检测场景**、**Bucket**、**分值**、**识别结果**、**Key**和**时间范围**自定义扫描范围。

   通过**导出**，将扫描结果导出进行查看。少于50条记录时将导出所有结果，否则只导出违规和疑似的结果。



> **说明** 通过单击扫描结果的图片或视频，可以查看**详细信息**，具体包括文件创建时间、Key值、所在Bucket。

如果扫描结果不符合您的业务需要，您可以对扫描结果进行自助审核，自助审核包含如下操作：

- **违规并删除**

  通过单击**违规并删除**，可将图片或视频从内容安全控制台和OSS Bucket中一并删除。支持单选或者多选。

- **正常并忽略**

  通过单击**正常并忽略**，则忽略该检测结果。忽略后该图片或视频将不再在控制台展示，并不影响存储在OSS Bucket中的图片或视频。支持单选或者多选。

- **正常并解冻**

  若您设置了自动冻结功能，则还可以在选中图片或视频后单击**正常并解冻**，将已冻结的图片或视频解冻。

您可以搜索您重点关注的结果，或者将扫描结果导出进行查看（少于50条记录时将导出所有结果，否则只导出违规和疑似的结果）。

后续步骤

您可以在**OSS违规检测 > 增量扫描**页面查看近7天的数据统计信息。更多信息，请参见查看统计数据。

1.5.3. 查看统计数据

OSS违规检测服务为您提供数据统计功能，当您完成增量扫描任务后，您可以随时在内容安全控制台查看数据统计信息。您可以通过监控一段时间的统计数据，根据网站内容的违规情况，对网站加强管控。

操作步骤

1. 登录内容安全控制台。
2. 在左侧导航栏，选择**OSS违规检测 > 增量扫描**。
3. 在**增量扫描**页面的操作列，选择**更多 > 数据统计**。
4. 在**数据统计**页面，通过单击**图片**、**视频**和**语音**页签查看最近7天扫描的统计信息。

   支持查看的数据统计信息如下表所示。

查询对象	支持的统计信息
图片	○ **图片总量**：表示检测的图片总数量。 ○ **鉴黄场景检测量**：包含违规、疑似、正常结果的数量。 ○ **暴恐涉政场景检测量**：包含违规、疑似、正常结果的数量。 ○ **广告场景检测量**：包含违规、疑似、正常结果的数量。 ○ **不良场景检测量**：包含违规、疑似、正常结果的数量。
视频	○ **视频总量**：表示检测的视频总数量。 ○ **截帧总量**：表示检测的视频截帧总数量。 ○ **鉴黄场景视频量**：包含违规、疑似、正常结果的数量。 ○ **暴恐涉政场景视频量**：包含违规、疑似、正常结果的数量。 ○ **广告场景检测量**：包含违规、疑似、正常结果的数量。 ○ **不良场景检测量**：包含违规、疑似、正常结果的数量。 ○ **语音反垃圾检测量**：包含违规、疑似、正常结果的数量。
语音	○ **语音总量**：包含违规、疑似、正常结果的数量。 ○ **语音总时长**：包含违规、疑似、正常结果的数量。 ○ **语音反垃圾检测量**：包含违规、疑似、正常结果的数量。

后续步骤

您可以在**OSS违规检测 > 增量扫描**页签查看近7天的扫描结果信息。更多信息，请参见查看扫描结果。

1.6. 存量扫描

1.6.1. 创建存量扫描任务

您可以手动创建一个存量扫描任务，对指定OSS Bucket中的已有图片或视频文件进行一次性违规检测，并在检测完成后查看扫描结果和数据统计信息。

前提条件

- 已经开通OSS对象存储服务并授权内容安全读取您的OSS Bucket。具体操作，请参见授权内容安全访问OSS存储空间。
- 已完成配置回调通知。具体操作，请参见配置消息通知。

背景信息

关于OSS违规检测的使用限制（例如支持的OSS Bucket、检测场景、文件格式等），请参见使用限制。

操作步骤

1. 登录内容安全控制台。
2. 在左侧导航栏，选择**OSS违规检测 > 存量扫描**。
3. 单击**创建扫描任务**，配置存量扫描任务。
4. 在**存量扫描任务-创建向导**对话框，完成以下配置。
   i. **选择Bucket**。

   从左侧**待选择**框，选中需要检测的Bucket，添加到右侧的**已选择**框。

   待选择框罗列了当前阿里云账号在OSS空间中可检测到的所有Bucket。

   然后单击**下一步**。

ii. **设置过滤条件**。

配置说明：

- **过滤或排除**：默认扫描已选Bucket中的所有文件，如果您需要指定扫描Bucket中的哪些文件或者不要扫描哪些文件，则可以单击Bucket右侧的**设置过滤条件**，并在**过滤条件**对话框，按照页面提示完成过滤条件配置，然后单击**确定**。

过滤条件参数说明

参数	说明
过滤条件	支持**包含**（表示扫描指定的路径）或**排除**（表示扫描时排除指定的路径）。 必须且只能选择一种方式。
路径	输入要包含或排除的路径。 单击**添加**可以添加多个路径。最多允许添加10个。 配置示例（以**包含**条件为例）： ■ 添加 `img/test_`，表示只扫描Bucket中以*img/test_*为前缀的文件。 ■ 如果要扫描的文件在特定目录下，您可以在文件名前加上目录路径，以整体作为前缀。例如，您要扫描的文件在*img/202011*目录下，且前缀为*test_*，您可以添加 `img/202011/test_` 作为路径。

- **时间范围**：必填。设置**文件上传时间**范围。内容安全只扫描在指定时间范围内上传的文件。

然后单击**下一步**。

iii. **配置检测场景。**

OSS违规检测为您提供8个推荐配置，每种推荐配置包含大类场景（图片、视频、语音）和细分场景（每种推荐配置的细分场景均不相同）。如果推荐的配置与您业务不符，您可以关闭或者开启大类场景，但不支持修改细分场景。

然后单击**下一步**。

iv. **配置扫描范围。**

扫描范围根据您配置的检测场景显示。例如，您在**场景配置**页面只勾选**图片**，那么**扫描配置**页面就只显示图片的配置信息。如果当前配置与您业务不符，您可以自定义扫描范围。

> **说明** 至少选择一类检测场景进行扫描，即图片、视频、语音中至少开启一项。

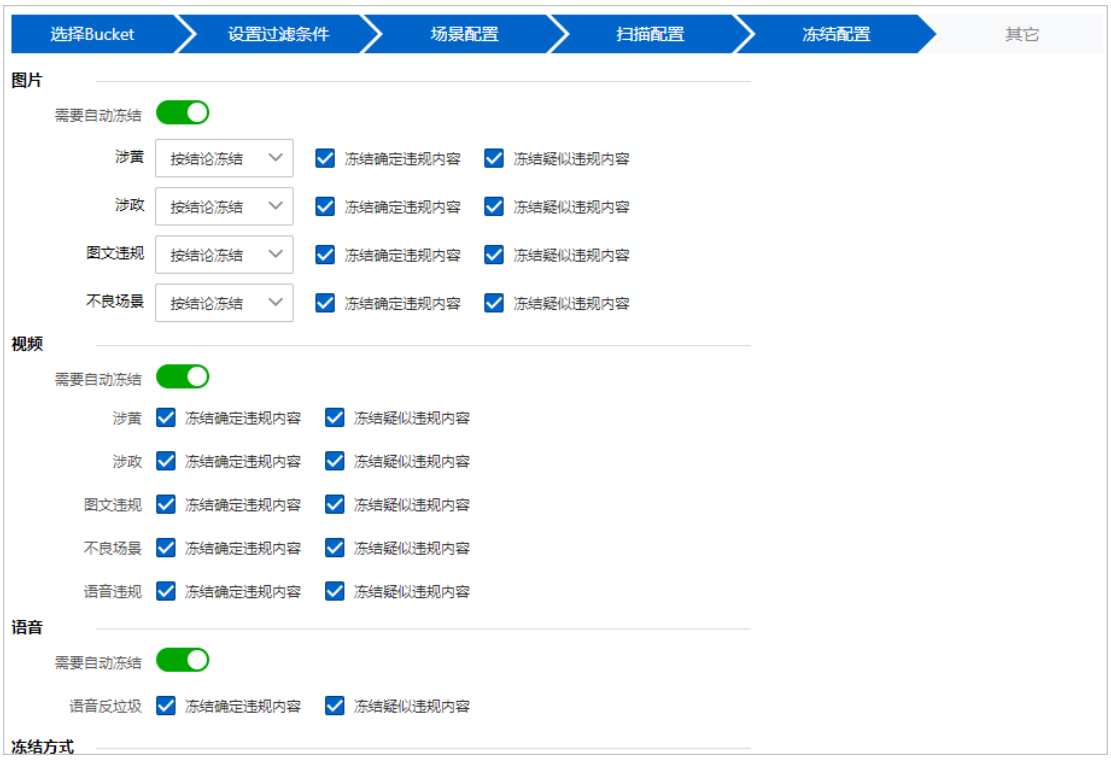
扫描配置参数说明

扫描范围	参数名称	说明
图片	**图片扫描上限**	扫描的图片张数默认为10000张。如果设置了扫描上限，扫描数量超出限制后将会停止扫描，因此会存在巨大违规图片外露的风险。常规情况下，不建议您设置扫描上限。
	检测无后缀文件	开启**检测无后缀文件**后，会扫描无后缀的图片文件，检测速度会降低，根据HTTP头的content-type判断是否为图片，支持content-type有：image/jpg，image/jpeg，image/png，image/gif，mage/bmp，image/webp。
视频	**视频扫描上限**	扫描的视频个数默认为1000个。如果设置了扫描上限，扫描数量超出限制后将会停止扫描，因此会存在巨大违规视频外露的风险。常规情况下，不建议您设置扫描上限。
	截帧频率	截帧频率默认为1帧/秒，您可以设置截帧频率（每多少秒截取一帧）。取值范围1~60帧/秒。截帧频率越高，识别准确率越高。
	单视频帧数上限	单视频帧数默认为200帧，您可以设置单个视频的最大截帧数量。取值范围：5~20000帧。
	单视频大小上限	单视频大小默认为500 MB，您可以设置单视频大小上限，取值范围：1~2048 MB，超过部分不会被检测。
语音	**语音扫描上限**	扫描的音频个数默认为1000个。如果设置了扫描上限，扫描数量超出限制后将会停止扫描，因此会存在巨大违规音频外露的风险。常规情况下，不建议您设置扫描上限。
	单音频大小上限	单音频大小默认为200 MB。您可以设置单音频大小上限，取值范围：1~2048 MB，超过部分不会被检测。

然后单击**下一步**。

v. **冻结配置**。

根据是否需要自动冻结检测出来的违规图片、视频和语音，配置对应的冻结规则，并选择一种**冻结方式**。



配置说明：

- 自动冻结违规图片：开启**图片**区域下的**需要自动冻结**开关，并为不同场景设置对应的冻结规则（不同场景的设置方法一样）。

  支持以下两种类型的冻结规则：

  - 按分值冻结：选择**按分值冻结**并设置一个违规分数阈值。当图片的违规分超过阈值时，将被冻结。

    > **注意** 请慎重修改冻结阈值。默认冻结分数是99.01分，正常情况下不建议设置到99分以下，分数过低可能会导致正常图片被冻结。

  - 按结论冻结：选择**按结论冻结**并选中要冻结的违规内容类型，可选项：**冻结确定违规内容**、**冻结疑似违规内容**。

    > **说明** 建议您优先使用按结论冻结，按分值冻结需具有一定的专业知识。如果您分值配置不适当可能会导致正常图片被冻结。

- 自动冻结违规视频：开启**视频**区域下的**需要自动冻结**开关，并为不同场景选择要冻结的违规内容类型，可选项：**冻结确定违规内容**、**冻结疑似违规内容**。
- 自动冻结违规语音：开启**语音**区域下的**需要自动冻结**开关，并为不同场景选择要冻结的违规内容类型，可选项：**冻结确定违规内容**、**冻结疑似违规内容**。

- 选择冻结方式：
  - **修改权限**：将您Bucket中public权限的违规文件设置为private访问权限。

    互联网用户无法读取private权限的文件，但您可以通过文件URL将私有文件分享给您的合作伙伴访问。更多信息，请参见在URL中包含签名。
  - **移动文件**：将您Bucket中违规的文件移动到Bucket中的备份目录（位置：*${bucket}/aligreen_freeze_backup/*），并删除原路径下的文件。

    > **注意** 选择该方式时请慎重。文件删除后将无法恢复，如需找回被删除的文件，您必须到备份目录中查找。

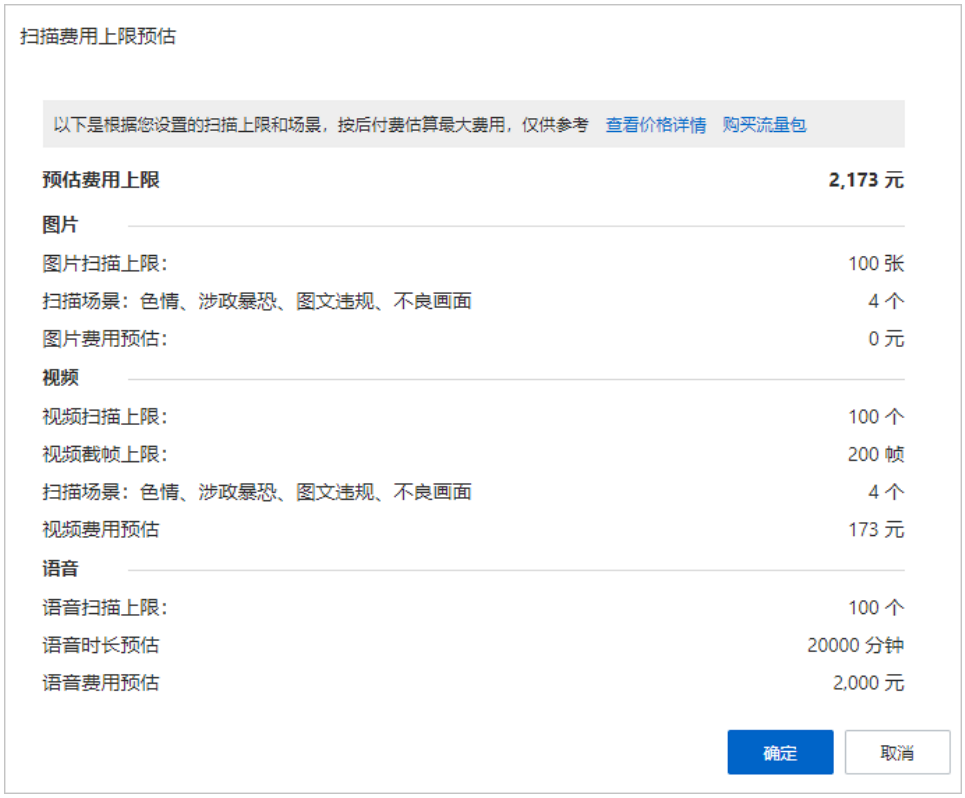
  然后单击**下一步**。

vi. （可选）**其它**。在**回调通知**区域，您可以选择一个已有的**回调通知方案**，通过指定的通知方案接收存量扫描的结果。

您必须先创建回调通知方案才可以进行设置。如果您未创建过回调通知方案，可以单击**新增通知方案**去添加通知方案。具体操作，请参见配置消息通知。

5. 在页面最下方，单击**提交**。

   OSS违规检测功能会根据您的配置为您预估出扫描费用的上限，您可以根据实际业务选择按量付费或者资源包抵扣方式。



然后单击**确定**。

存量扫描任务提交后，将立即开始扫描。扫描所需时间取决于存量文件的数量和开启的扫描场景。您可以在存量扫描任务列表中查看当前任务的状态。

OSS违规检测为您提供如下功能。您需要根据实际需要，选择合适的操作。

- **暂停扫描**

  如果您因为其他原因，需要暂停扫描任务时，单击操作列**暂停扫描**。只有扫描任务未结束时可以执行该操作，支持扫描任务暂停后7天内可以继续扫描。

- **数据统计**

  如果您需要查看当前任务的调用量统计信息，单击操作列**数据统计**，进入**OSS违规检测调用量**页面，查看最近7天的调用量。您可以设置扫描的时间、任务等条件自定义查询统计信息。

  更多内容，请参见查看统计数据。

- **终止扫描**

  如果您不再需要扫描当前任务时，单击操作列**终止扫描**，终止扫描后存量的文件将不再执行扫描动作。只有扫描任务未结束时可以执行该操作。

- **扫描结果**

如果您需要查看当前任务的扫描结果，在操作列选择**更多 > 扫描结果**（任务未扫描完成）或者单击操作列**扫描结果**（任务已扫描完成）。您可以设置扫描的时间、任务等条件自定义查询结果信息。

更多内容，请参见查看扫描结果。

- **查看配置**

  如果您需要查看当前任务的配置信息，在操作列选择**更多 > 查看配置**（任务未扫描完成）或者单击操作列**查看配置**（任务已扫描完成），展开**任务配置详情**面板查看。

- **用量说明**

  如果您需要查看当前任务的用量，在操作列选择**更多 > 用量说明**，展开**任务用量说明**面板查看。

后续步骤

您可以在**OSS违规检测 > 存量扫描**页面查看近7天的扫描结果和数据统计信息。更多信息，请参见查看统计数据和查看扫描结果。

1.6.2. 查看扫描结果

OSS违规检测服务为您提供查看扫描结果的功能，当您完成增量扫描任务后，您可以随时在内容安全控制台查看扫描结果，并根据检测结果执行自助审核。

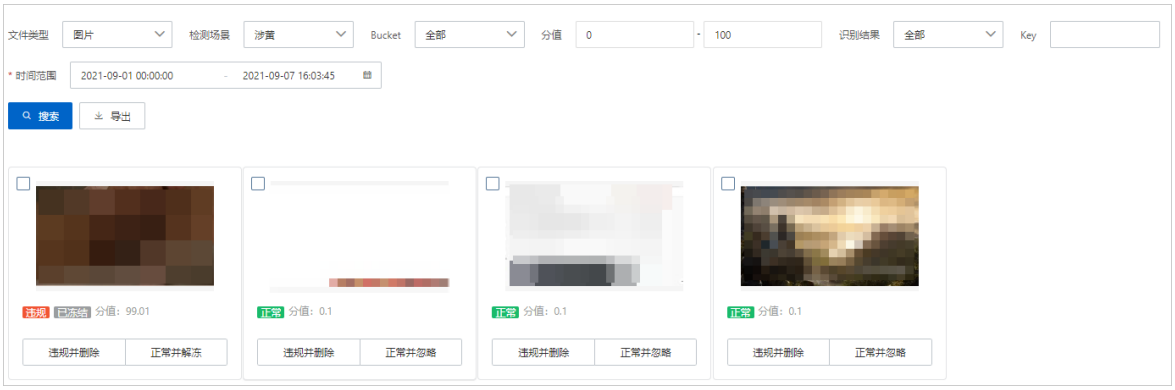
操作步骤

1. 登录内容安全控制台。
2. 在左侧导航栏，选择**OSS违规检测 > 存量扫描**。
3. 在**存量扫描**页面，查看您的**任务概览**、**任务状态**以及**任务结果**。

4. 单击操作列**扫描结果**，查询扫描结果并进行自助审核。

   默认显示最近7天的扫描结果和处理的违规内容。您可以设置**文件类型**、**检测场景**、**Bucket**、**分值**、**识别结果**、**Key**和**时间范围**自定义扫描范围。

   通过**导出**，将扫描结果导出进行查看。少于50条记录时将导出所有结果，否则只导出违规和疑似的结果。



> **说明** 通过单击扫描结果的图片或视频，可以查看**详细信息**，具体包括文件创建时间、Key值、所在Bucket。

如果扫描结果不符合您的业务需要，您可以对扫描结果进行自助审核，自助审核包含如下操作：

- **违规并删除**

  通过单击**违规并删除**，可将图片或视频从内容安全控制台和OSS Bucket中一并删除。支持单选或者多选。

- **正常并忽略**

  通过单击**正常并忽略**，则忽略该检测结果。忽略后该图片或视频将不再在控制台展示，并不影响存储在OSS Bucket中的图片或视频。支持单选或者多选。

- **正常并解冻**

  若您设置了自动冻结功能，则还可以在选中图片或视频后单击**正常并解冻**，将已冻结的图片或视频解冻。

您可以搜索您重点关注的结果，或者将扫描结果导出进行查看（少于50条记录时将导出所有结果，否则只导出违规和疑似的结果）。

后续步骤

您可以在**OSS违规检测 > 存量扫描**页面查看近7天的数据统计信息。更多信息，请参见查看统计数据。

1.6.3. 查看统计数据

OSS违规检测服务为您提供数据统计功能，当您完成存量扫描任务后，您可以随时在内容安全控制台查看数据统计信息。您可以通过监控一段时间的统计数据，根据网站的之前数据的违规情况，对网站的存量内容加以调整。

操作步骤

1. 登录内容安全控制台。
2. 在左侧导航栏，选择**OSS违规检测 > 存量扫描**。
3. 在**存量扫描**页面的**操作**列，单击**数据统计**。
4. 在**数据统计**页面，通过单击**图片**、**视频**和**语音**页签查看最近7天扫描的统计信息。

   支持查看的数据统计信息如下表所示。

查询对象	支持的统计信息
图片	○ **图片总量**：表示检测的图片总数量。 ○ **鉴黄场景检测量**：包含违规、疑似、正常结果的数量。 ○ **暴恐涉政场景检测量**：包含违规、疑似、正常结果的数量。 ○ **广告场景检测量**：包含违规、疑似、正常结果的数量。 ○ **不良场景检测量**：包含违规、疑似、正常结果的数量。
视频	○ **视频总量**：表示检测的视频总数量。 ○ **截帧总量**：表示检测的视频截帧总数量。 ○ **鉴黄场景视频量**：包含违规、疑似、正常结果的数量。 ○ **暴恐涉政场景视频量**：包含违规、疑似、正常结果的数量。 ○ **广告场景检测量**：包含违规、疑似、正常结果的数量。 ○ **不良场景检测量**：包含违规、疑似、正常结果的数量。 ○ **语音反垃圾检测量**：包含违规、疑似、正常结果的数量。

查询对象	支持的统计信息
语音	○ **语音总量**：包含违规、疑似、正常结果的数量。 ○ **语音总时长**：包含违规、疑似、正常结果的数量。 ○ **语音反垃圾检测量**：包含违规、疑似、正常结果的数量。

后续步骤

您可以在**OSS违规检测 > 存量扫描**页面查看近7天的扫描结果。更多信息，请参见查看扫描结果。

2.机器审核

2.1. 自定义机审标准

内容安全采用阿里云默认的机器审核标准为您提供内容检测服务。如果您在使用过程中发现默认的审核标准对您的业务需求过于严格或者宽松，您可以使用内容安全的自定义机审标准功能。本文介绍了如何配置自定义机审标准。

背景信息

业务场景（BizType）：审核标准基于业务场景配置，每个业务场景对应一套审核标准。未配置自定义审核标准时，统一使用默认的业务场景以及对应的审核标准。配置自定义业务场景后，您必须在内容检测API的接口中传递自定义业务场景，检测才会按照自定义业务场景的标准进行。

操作步骤

1. 登录内容安全控制台。
2. 在左侧导航栏，选择**设置 > 机器审核**。
3. 创建业务场景。

   如果您已创建业务场景，请跳过该步骤。

   i. 在**机器审核**页面的**业务场景管理**页签，单击**创建业务场景**。

   ii. 在**创建bizType**对话框中，请参考如下表格设置业务场景名称、行业模板、行业分类等信息。

参数	说明
BizType名称	为业务场景命名。支持使用数字、英文字符、下划线（_），且不超过32个字符。
行业模板配置	具有行业标准的策略配置。使用行业模板的情况下，部分机审标准功能不可自定义，如不使用，则完全由您自行配置。
行业分类	业务所属行业分类。若传入行业分类，我们能够更好地帮助您调整策略配置。
从现有导入	如果您已经创建过业务场景，您可以直接导入已创建的业务场景的配置。
描述	对业务场景的补充说明。支持使用中英文、数字、下划线组合，不超过32个字符。

   iii. 单击**确定**，成功创建业务场景。

4. 设置业务场景。

   您可以配置**机审标准**、**入审数据**、**证据转存**。

   - **机审标准**

     不同检测场景的审核标准定义不同，具体以控制台显示为准。

     a. 在**机器审核**页面，单击目标业务场景右侧操作列的**编辑**。

b. 根据实际情况选择您需要设置的机审标准。

目前内容安全支持设置图片、视频、文本和语音的机审标准。其中，图片和视频的机审内容包括色情、涉证暴恐、不良画面和图文违规；文本和语音机审内容主要是指文本或者语音是否涉黄、涉政、辱骂、广告和违禁。

如果您的业务中不需要某个检测场景，那么不设置对应的检测场景即可。例如，您不需要检测广告，那么在设置机审标准中不配置广告相关的内容即可。

说明 当您设置图片的机审标准时，您可以通过页面右侧的**关联图库**和**关联文本库**关联风险库中的配置信息。

c. 单击**保存**，完成自定义机审标准。

在调用接口时，您需要传递对应的业务场景，设置的自定义机审标准才会生效。例如，检测图片涉黄风险时（请参见同步检测），请求参数scene需要传递*porn*，标准才会生效。

- **入审数据**

a. 单击**入审数据**。

b. 在**入审数据**页签，勾选需要流入自助审核页面的数据类型。

关于自助审核的具体操作，请参见自助审核。

- **证据转存**

使用内容检测API时，若您希望保存机审识别的结果证据，您可以开启证据转存功能。证据转存与阿里云对象存储OSS配合使用，支持将视频、语音、图片检测中识别到的违规、疑似、正常内容转存到您指定的OSS存储空间，并返回转存文件的OSS URL链接。本文介绍了开启和配置证据转存的方法。

证据转存目前只支持视频、语音、图片对象的检测。本文所述转存文件特指机审结果为违规（返回suggestion=block）、疑似（返回suggestion=review）或正常（返回suggestion=pass）的被检测视频、语音或者图片。

证据转存默认关闭，若需使用，您必须开启并配置视频证据转存、语音证据转存或者图片证据转存。

- 开启视频证据转存后，机审结果违规、疑似、正常的视频文件（含视频流）和视频截帧图片会转存到您配置的Bucket中。
- 开启语音证据转存后，机审结果违规、疑似、正常的语音文件（含语音流）和语音分片会转存到您配置的Bucket中。
- 开启图片证据转存后，机审结果违规、疑似、正常的图片文件会转存到您配置的Bucket中。

a. 单击**证据转存**。

b. 在**证据转存**页签，根据需要开启**视频证据转存**、**语音证据转存**或**图片证据转存**，完成相关配置。

> **说明** 如果您是初次登录，则需要根据页面提示完成OSS授权，授予内容安全对对象存储服务的读写权限（复用当前OSS违规检测的授权）。如果已经完成授权，则直接进入配置页面。

配置项	说明
转存Bucket	用来存储证据文件的OSS存储空间。
转存目录	用户传递目录。所有证据文件按照转存规则存放在指定的用户传递目录下。更多信息，请参见转存规则说明。 **说明** 若指定的目录在转存Bucket中不存在，则会自动创建。
访问链接有效期	证据文件转存后生成的访问链接的有效期，取值范围：300～3600（秒）。
转存范围	目前支持以下三种转存范围： ■ **机器识别违规**：转存机器识别的违规内容。 ■ **机器识别疑似**：转存机器识别的疑似违规内容。 ■ **机器识别正常**：转存机器识别的正常内容。

c. 单击**保存**。

成功配置证据转存后，在下次调用视频审核、语音反垃圾检测、图片审核并检测出违规、可疑或者正常的视频、语音或图片内容时，您将在返回参数（视频审核对应data.extras.newUrl、语音反垃圾对应data.new_url和data.result.details.url和图片审核对应data.storedUrl）中查看转存文件的OSS URL链接。您也可以前往OSS控制台，在转存Bucket中查看转存文件。

转存规则说明

视频转存规则

文件类型	存放目录	命名规则
视频文件	*${bucket}/${用户传递目录}/video/${suggestion}/${taskId}/${视频名称.后缀}*	转存的视频文件以原视频文件名称命名。
视频截帧	*${bucket}/${用户传递目录}/video/${suggestion}/${taskId}/frames/${截帧图片名.后缀}*	转存的截帧图片以截帧的时间点命名。例如*00_01_02*，表示00时01分02秒的截帧。
视频流	*${bucket}/${用户传递目录}/video/${suggestion}/${taskId}/${视频流名称.后缀}*	转存的视频流以开始检测时间命名。例如*20190102_12_02_03.wav*，表示2019年01月02号12时02分03秒开始检测的视频流。

语音转存规则

文件类型	存放目录	命名规则
语音文件	*${bucket}/${用户传递目录}/audio/${suggestion}/${taskId}/${语音名称.后缀}*	转存的语音文件以原语音文件名称命名。
语音分片	*${bucket}/${用户传递目录}/audio/${suggestion}/${taskId}/slice/${语音分片名.后缀}*	转存的语音分片以语音开始和结束时间命名。例如*00_01_02-00_10_13.mp3*，表示语音中从00时01分02秒到00时10分13秒的语音片段。

图片转存规则

文件类型	存放目录	命名规则
图片文件	*${bucket}/${用户传递目录}/image/${suggestion}/${taskId}/${图片名称.后缀}*	转存的图片文件以原图片文件名称命名。

2.2. 配置消息通知

内容安全支持以异步消息通知的方式向您发送内容识别和您自助审核的结果。如果您的业务需要使用或集成相关数据，则可以配置回调通知。本文介绍了配置内容检测API回调通知的操作方法。

背景信息

> **说明** 内容检测API和OSS违规检测均支持回调通知。关于OSS违规检测回调通知，请参见OSS违规检测回调通知。

内容检测API回调通知分为扫描结果回调通知和审核结果回调通知。

- 扫描结果回调通知：内容安全完成检测请求后，服务端将识别的结果以POST请求的方式，发送到您设置的HTTP回调通知地址。
- 审核结果回调通知：您通过自助审核或调用反馈接口修改识别的结果后，内容安全服务端将审核结果以POST请求的

方式，发送到您设置的HTTP回调通知地址。更多信息，请参见自助审核。

相关概念

在配置回调通知前，请了解下表中描述的相关概念。

名称	说明
回调地址	回调地址是您在内容安全控制台配置的服务端地址，通常是您自己的业务服务器的公网地址。回调地址需要满足以下要求： • 应为HTTP、HTTPS协议接口的公网可访问的URL。 • 支持POST方法。 • 支持传输数据编码采用UTF-8。 • 支持数据接收格式为 `application` 、 `x-www-form-urlencoded` 。 • 支持表单参数checksum和content。
Seed	Seed值用于校验发送到您设置的HTTP回调通知地址的请求是否来自内容安全服务端。
回调次数	您的服务端接收到内容安全推送的回调消息后，如果返回的HTTP状态码为200，表示接收成功；如果返回其他的HTTP状态码，均视为接收失败。如果接收失败，内容安全服务端会重复推送回调消息，直至您接收成功。内容安全最多重复推送回调消息16次。
回调数据	回调数据是内容安全服务端向您设置的回调通知地址返回的数据内容。回调数据的结构描述见回调通知表单数据。

回调通知表单数据

名称	类型	描述
checksum	String	总和校验码，由 `<用户uid> + <seed> + <content>` 拼成字符串，通过SHA256算法生成。用户UID即阿里云账号ID，可以在阿里云控制台上查询。 **说明** 为防篡改，您可以在获取到推送结果时，按上述算法生成字符串，与checksum做一次校验。
content	String	字符串格式保存的JSON对象，请自行解析反转成JSON对象。关于content解析成JSON后的结构，请参见下文content说明。 **说明** 在内容检测API和OSS违规检测中，content的参数结构不同。

扫描结果回调通知

内容检测API的异步检测接口均支持扫描结果回调，例如图片审核异步检测、视频审核异步检测。您在发起异步调用时，如果希望内容安全向您返回扫描结果，则必须在请求参数中传入callback（回调地址）和seed（校验字符串）。

说明 在异步调用时如果不使用扫描结果回调，则您只能通过定期轮询的方式获取异步检测结果。

操作步骤

1. 自行准备好接收扫描结果的HTTP回调地址和Seed参数。
2. 调用内容检测异步API接口时，传递相应的callback和seed请求参数，具体请参见API接口描述中的参数说明。

人工审核回调通知

人工审核接口不支持实时返回检测结果，系统支持回调方式通知调用方。人工审核分为人机审核和纯人工审核，两者的回调方式有一定的差异，以下分别进行介绍。

- **人机审核**

  如果您使用的是阿里云人工审核服务+机审服务，人工审核模块的回调需要通过控制台配置消息通知来实现，具体操作步骤如下：

  i. 登录内容安全控制台。
  ii. 在左侧导航栏，选择**设置 > 机器审核**。
  iii. 在**机器审核**页面，单击**消息通知**。
  iv. 在**消息通知**页签，单击**新增通知方案**。
  v. 在**新增通知方案**对话框，输入**方案名称**、**回调地址**，选择**加密算法**、**通知类型**、**审核结果**。然后单击**确认**。

  保存完成后，系统自动生成seed。seed值用于校验您的回调接口收到的请求来自阿里云。请保存自动生成的seed。

  > 注意
  > - 如果您已经配置了机审回调通知服务，您可以直接复用之前的配置，也可以根据业务需要重新配置。
  > - 系统支持对于所有的审核结果进行回调，您也可以根据自己的业务场景选择部分审核结果进行回调。
  > - 通知类型项请选择**阿里云人审结果**。
  > - 加密算法方式：
  >   - SHA256：使用HMAC-SHA256加密算法。
  >   - 国密SM3：使用国密SM3加密算法，返回十六进制的字符串，且字符串由小写字母和数字组成。
  >
  >   例如，abc经国密SM3加密后返回66c7f0f462eeedd9d1f2d46bdc10e4e24167c4875cf2f7a2297da02b8f4ba8e0。

  vi. 在**机器审核**页面，单击**业务场景管理**。
  vii. 在**业务场景管理**页签，单击目标业务场景右侧**操作列**的**关联消息通知**，关联您配置的**回调通知方案**。

- **纯人工审核**

  如果您使用的是阿里云纯人工审核服务，人工审核模块的回调需要通过请求参数传入callback（回调地址）和seed（校验字符串）来实现。

  i. 自行准备好接收扫描结果的HTTP回调地址和seed参数。
  ii. 调用内容检测异步API接口时，传递相应的callback和seed请求参数。

  关于具体参数说明，请参见图片人工审核、视频人工审核、文本人工审核、复核人工审核结果。

content说明

启用回调通知后，内容安全将按照回调配置发送内容检测API回调通知。回调通知中包含content表单数据。下表描述了content表单字段的结构。

content表单字段结构描述

名称	类型	是否必须	说明

名称	类型	是否必须	说明
scanResult	JSONObject	否	扫描结果。根据检测对象（图片、视频）的不同，结构有差异。 • 针对图片对象，结构与图片同步检测中的results返回参数一致，具体请参见图片同步检测。 • 针对视频对象，结构与视频异步检测中的results返回参数一致，具体请参见视频异步检测。
auditResult	JSONObject	否	您的自助审核结果。发生自助审核操作时才会有该字段。具体请参见auditResult。 说明 如果只推送扫描结果，则没有该字段。
humanAuditResult	JSONObject	否	阿里云的人工审核结果。如果您购买了阿里云的人工审核服务，则人工审核的结果体现在该字段中。具体请参见humanAuditResult。

auditResult

名称	类型	是否必须	说明
suggestion	String	是	自助审核的结果，取值： • *block*：审核时设置违规。 • *pass*：审核时设置正常。
labels	JSONArray	否	自助审核时设置的标签，包含以下可选值中的一个或者多个： • *porn*：鉴黄。 • *terrorism*：暴恐涉政。 • *ad*：图文广告。 • *live*：不良场景。

humanAuditResult

名称	类型	是否必须	说明
suggestion	String	是	阿里云人工审核的结果，取值： • *block*：阿里云人工审核结果为违规。 • *pass*：阿里云人工审核结果为正常。
taskId	String	是	检测任务的ID。通过任务ID可以关联到对应内容的审核结果。
dataId	String	是	检测内容的ID。
labels	StringArray	否	人工审核的标签结果，可能有多个值。 说明 默认不返回该参数。配置该字段需要额外收费，具体请联系商务沟通。

content示例

```
{
    "scanResult": {
        "code": 200,
        "msg": "OK",
        "taskId": "fdd25f95-4892-4d6b-aca9-7939bc6e9baa-1486198766695",
        "url": "http://1.jpg",
        "results": [
            {
                "rate": 100,
                "scene": "porn",
                "suggestion": "block",
                "label": "porn"
            }
        ]
    },
    "auditResult": {
        "suggestion": "block",
        "labels": [
            "porn",
            "ad",
            "terrorism"
        ]
    },
    "humanAuditResult": {
          "suggestion": "pass",
          "dataId": "yyyy",
          "labels": [
              "色情",
              "低俗"
        ],
          "taskId": "xxxxxx"
    }
}
```


2.3. 自定义OCR模板

内容检测API的OCR卡证内容识别功能支持自定义OCR模板，帮助您提取自定义图片中的结构化文字信息。您可以在内容安全控制台自定义OCR模板，根据需要配置要识别的图片模板和待识别的文字信息，实现对各种类型的票据、证件等图片进行文字识别。

背景信息

如果您需要识别的图片类型不在已有的结构化OCR支持范围内，您可以使用自定义OCR模板。

进行本文操作前，请先熟悉以下概念。

- 模板：格式和包含信息完全相同的一类图片生成的一种规范版式。进行图片文字识别前，您需要在内容安全控制台手动创建模板。创建成功的模板将获得一个唯一的ID作为其标识。在调用OCR检测接口时，您需要传入要应用的模板ID作为请求参数。

  创建模板时，您需要上传一张待识别的图片作为样本。样本图片需要满足以下要求：

  - 使用.png、.jpg、.jepg、.bmp、.gif格式。
  - 大小在1 KB到10 MB之间，分辨率在320*320像素到4096*4096像素之间。
  - 尽量摆放端正平整，不存在模糊、过度曝光、阴影等不良情况。
  - 尽量突出需要识别的部分。建议您手动剪裁掉不需要部分，以提高识别准确率。
  - 至少存在四个模板参考字段，且尽量分散在图片的边缘（越分散越好），用于准确定位模板。

- 选取的模板参考字段、待识别字段的高度不小于20像素。

- 参考字段：用于定位模板位置的固定字段。参考字段的选取会影响图片的识别准确率。参考字段务必选取位置和内容都不会变化的文字内容。单个参考字段内的文字不可以换行，建议您选取四个以上的参考字段。
- 识别字段：需要识别的内容字段。设置识别字段时，需要给字段设置key值，最终识别结果会以`key:识别内容`格式返回。

操作步骤

1. 登录内容安全控制台。
2. 在左侧导航栏，选择**设置 > 机器审核**。
3. 在**机器审核**页面，单击**OCR模板**。
4. 单击**创建模板**。

5. 在**创建模板**面板，设置**模板名称**，并单击**选择图片**，选择一张待识别的图片作为样本上传。

成功创建模板。新建的模板显示在左侧**模板列表**中，选择新建的模板进行后续的模板配置。

6. 设置参考字段。
   i. 单击**设置参考字段**。

ii. 在**模板编辑**页面，单击**新增字段**并用绿色矩形框框选图片上位置固定不变的单行参考字段。

> **说明** 参考字段区域务必框选单行文字，且尽量将文字包裹完整。

iii. 重复上一步，设置至少四个不同的参考文字区域后，单击**保存**。

7. 设置识别字段。
   i. 单击**设置识别字段**。
   ii. 在**模板编辑**页面，单击**新增字段**并用绿色矩形框框选待识别的单行文字，为框选中的内容设置一个Key值，作为识别结果的标识。

   > **说明** 如果要识别的字段分多行显示，建议您分别框选单行文字，并为它们设置相同的Key值（例如，下图中的date键值分两行显示，分别框选它们并设置两个名称为date的Key）。算法会将多行Key值相同的字段以框选顺序组合返回。

   iii. 重复上一步，添加完所有待识别内容，并单击**保存**。
8. 完成模板创建后，选择要应用的模板，单击**复制模板ID**。
9. 参见OCR同步检测，调用检测接口进行图片OCR识别，并将复制的模板ID作为请求参数extras结构体中的templateId传入，应用自定义OCR模板。

2.4. 风险库管理

2.4.1. 自定义图库

内容安全默认依据阿里巴巴全局风险图库为您提供检测服务，可以满足大部分的常规检测需求。为了使检测结果更贴合您的实际业务，内容安全也支持自定义图库。您可以使用自定义图库来管理需要针对性地拦截、放行、人工审核的图片，应对突发的管控需求。

背景信息

根据用途不同，自定义图库分为黑名单、白名单、疑似名单。在检测中应用自定义图库后，若被检测图片命中图库中的样本，则会被打上图库对应的识别结果标签。黑名单图库对应的识别结果是违规（拦截），白名单对应正常（放行），疑似名单则对应疑似（人工审核）。自定义图库包括系统回流图库和用户创建图库。

- 系统回流图库由您的自助审核记录自动生成，默认应用于所有同类场景的图片检测。您可以管理系统回流图库中的图片，不可以操作系统回流图库，例如停用或删除图库。
- 用户创建图库由您自行添加，可用于某次检测或某类检测场景。您可以管理用户创建图库中的图片，也可以操作用户创建图库。

> **说明** 您可以创建10个自定义图库（不含系统回流图库），且在每个图库中添加最多10000张样本图片。

内容检测API、OSS违规检测、站点检测均支持自定义图库，且内容检测API和OSS违规检测的自定义图库互通。在内容检测API中，自定义图库适用于以下场景：图片/视频鉴黄、图片/视频暴恐涉政识别、图片/视频广告识别、图片/视频不良场景识别。

下文介绍了在内容安全控制台管理内容检测API自定义图库的操作方法。除了控制台操作，您还可以通过API接口或SDK完成相关操作。更多内容，请参见：

- 使用API管理自定义图库
- 使用Java SDK管理自定义图库

操作步骤

1. 登录内容安全控制台。
2. 在左侧导航栏，选择**机器审核 > 风险库管理**。
3. 在**风险库管理**页面，单击**创建图库**。
4. 在**创建图库**对话框中，请参考下表完成图库配置，并单击**确定**。

   创建图库参数说明表

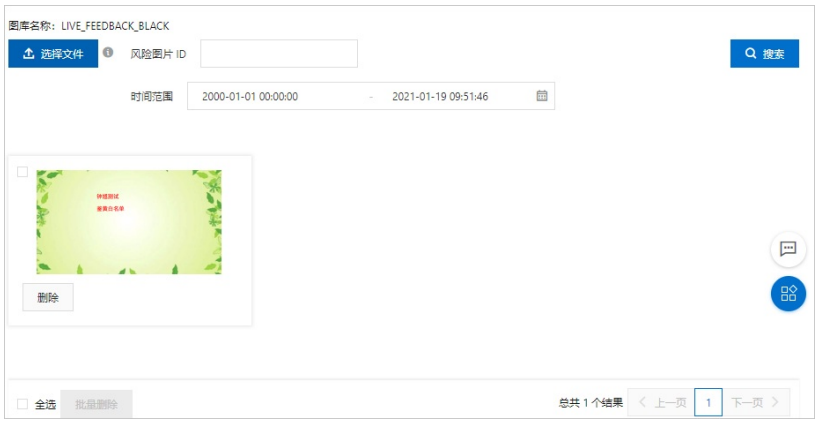
配置项	说明
名称	为图库命名。名称不超过64个字符。建议设置为可读性较强的中文名称。
使用场景	选择图库的使用场景，取值： ○ **鉴黄**：图片、视频鉴黄（scenes包含porn）。 ○ **暴恐**：图片、视频暴恐涉政识别（scenes包含terrorism）。 ○ **广告**：图片、视频图文违规识别（scenes包含ad）。 ○ **不良场景**：图片、视频不良场景识别（scenes包含live）。
识别结果	选择图库的用途，取值： ○ **黑名单**：若命中图库中样本，则机审结果返回违规。 ○ **疑似名单**：若命中图库中样本，则机审结果返回疑似。 ○ **白名单**：若命中图库中样本，则机审结果返回正常。
BizType	BizType允许您根据不同的业务需求配置并应用不同的图库，例如使用BizType指定在某次检测中应用图库。BizType生效逻辑如下： ○ 图库设置BizType为“A”，检测时API中传递了BizType为“A”，则检测时只会使用BizType为“A”的图库（前提是图库已启用）。 ○ 其他情况下，检测均会使用所有已启用的图库。 **说明** 包含Biztype属性的风险图库不适用于OSS违规检测。

   成功创建图库，已创建图库默认启用。

5. 在**自定义图库**页签，单击指定图库右侧操作列下的**管理**。

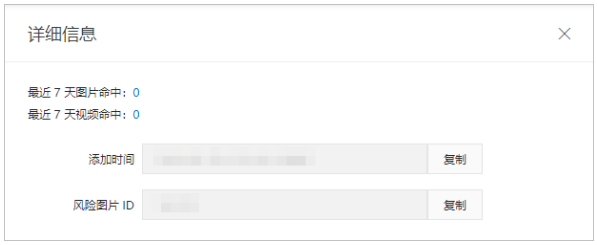
   自定义图库页签下显示所有自定义图库，其中系统回流图库的名称前有**系统**标识，且按照使用场景_FEEDBACK_WHITE/BLACK命名。例如，**LIVE_FEEDBACK_BLACK**是由系统回流生成的用于不良场景的黑名单图库。

6. 在**图库管理**页面，您可以根据需要执行如下操作。



> **说明** 在图库中新增、删除图片样本，大约需要15分钟生效。

- 根据**风险图片ID**、**时间范围**查询图片。
- 单击图片，展开图片的**详细信息**面板，查看近期命中数量、添加时间、风险图片ID等信息。



- 单击**选择文件**，上传图片到图库。

> **说明** 支持上传PNG、JPG、JEPG和BMP格式的图片。

- 单击图片下的**删除**，删除图片；选中多张图片后，单击**批量删除**，批量删除图片。

相关操作

对于自定义图库（非系统回流图库），您可通过**自定义图库**页签中**操作**列下的**删除**、**编辑**和**停用**对目标图库进行操作。

2.4.2. 自定义文本库

为了使检测结果更贴合您的实际业务，内容安全支持自定义文本库。自定义文本库适用于图片审核（图文违规识别）、文本反垃圾、文件反垃圾、语音反垃圾场景。您可以使用自定义文本库来管理需要针对性地拦截、放行（忽略）、人工审核的文本，应对突发的管控需求。

背景信息

> **注意** 建议您在使用自定义文本库前，仔细阅读本文档，了解文本库使用方式。避免因关键词添加不当导致误抓，影响检测效果。

自定义文本库包括系统回流文本库和用户创建文本库。

- 系统回流文本库由您的自助审核记录自动生成，默认应用于所有同类场景的检测。您可以管理系统回流文本库中的文本，但是不可以对系统回流文本库进行停用或删除操作。关于自助审核，请参见自助审核。
- 用户创建文本库由您自行添加，可用于某次检测或某类检测场景。您可以管理用户创建文本库中的文本，也可以操作用户创建文本库。

> **说明** 您可以创建10个自定义文本库（不含系统回流文本库）。

内容检测API、站点检测均支持自定义文本库。在内容检测API中，自定义文本库适用于以下场景：图片审核（图文违规识别）、文本反垃圾、文件反垃圾、语音反垃圾。

下文介绍了在内容安全控制台管理内容检测API自定义文本库的操作方法。除了控制台操作，您还可以通过API接口或SDK完成相关操作，具体请参见以下文档：

- 使用API管理自定义文本库
- 使用Java SDK管理自定义文本库

文本类型

自定义文本库的文本类型包括关键词和相似文本。

- 关键词

  关键词是针对短小词语进行防控的一种方式。您可以将其理解为：一句话或者一段文本里面是否包含某个既定词语，当包含该词语时，则表明命中该关键词。不同的业务场景支持配置不同的关键词。

  在内容安全的识别中，关键词技术可以被应用到图文违规、文本反垃圾、语音反垃圾场景中，具体配置见对应场景中的使用描述（配置参数可能略有出入）。

  中文关键词支持“与（&）”、“非（~）”的逻辑判断属性。示例：

  - 定义“A&B”，则只有在句子中同时出现“A”和“B”时，才会命中。
  - 定义“A~B”，则只有在句子中只出现“A”且不出现“B”时才会命中，同时出现“A”和“B”则不会命中。

  > **说明** 与（&）必须在非（~）之前。例如，您可以设置“A&B~C”作为关键词，但不能设置“A~C&B”作为关键词。

- 相似文本

  相似文本是针对句子或者段落式文本进行相似性判断的一种方式。您可以将其理解为：两句话或者两段文本，从句意上具有非常强的相似性，但又不是百分百一样，局部可能有变化，整体上却具有相同的意思或者在描述同一件事情。通过既定或者参照的文本样本，可以判断要识别的文本是否与样本具有强相似性。当相似性的概率在一定程度上时，则表明命中样本。

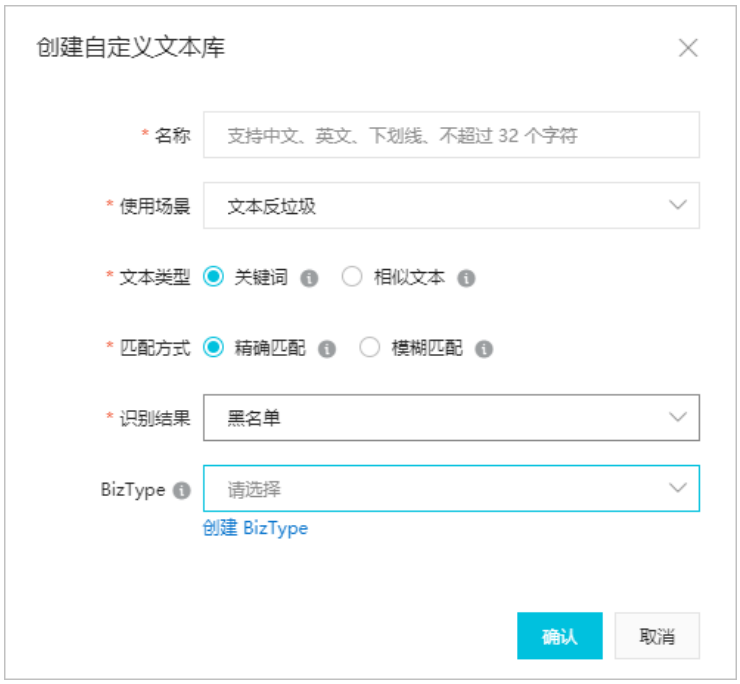
  相似文本文本库适用于文本反垃圾的检测场景。通过定义自己业务的相似文本库黑名单、白名单、疑似名单（疑似名单是指业务上需要识别出来，且需要人工审核），并在相似文本库里面维护与您业务相关的文本样本，从而指导文本反垃圾识别去过滤命中相似文本样本的内容。

使用限制

类型	项目	限制
文本库	库个数	不超过10个（不含系统回流文本库）。
文本库	库名长度	不超过20个字符。
关键词	关键词类型	• 支持中文关键词。 • 支持用字母和数字作为关键词。 说明 检测时字母和数字会被当作整体进行分词。 • 暂不支持英文关键词。
关键词	单个文本库中关键词个数	不超过10000个。
关键词	关键词最大长度	50个字符（包括符号）。
关键词	中文关键词编码类型	UTF-8。
关键词	关键词格式	不允许包含以下特殊字符（包括全角）： @ # $ % ^ * ( ) < > / ?, . ; _ + - = ‘ “ 空格 tab键
相似文本	相似文本长度	10~4000个字符。 说明 如果添加的文本过长，容易引起文本误抓。建议文本长度不要超过200个字符。
相似文本	单个文本库中相似文本个数	不超过10000个。
相似文本	文本编码格式	UTF-8。
相似文本	相似文本内容	文本样本需要包含明确的可提取的中文语义特征。如果经过引擎分析特征数太少，该文本样本将不会生效，引擎将其直接忽略。 说明 如果一段样本都是无意义的字母数字，或各种表情符等，则可能被忽略。

操作步骤

1. 登录内容安全控制台。
2. 在左侧导航栏，选择**机器审核 > 风险库管理**。
3. 单击**创建文本库**。
4. 在**创建自定义文本库**对话框，情参考创建自定义文本库参数说明表完成文本库配置，并单击**确认**。



创建自定义文本库参数说明表

参数	说明
名称	为文本库命名。文本库名称允许重复，但建议您在业务中将其设置为唯一。
使用场景	选择文本库的使用场景。取值： ○ **文本反垃圾**：文本反垃圾（scenes包含antispam）、文件反垃圾（textScenes包含antispam） ○ **语音反垃圾**：语音反垃圾（scenes包含antispam） ○ **图片广告**：图片审核（scenes包含ad）
文本类型	选择文本库的文本类型。取值： ○ **关键词**：使用关键词匹配，只要包含关键词就会命中，覆盖面大。 ○ **相似文本**：使用文本相似度匹配，只有整段文本相似才会命中，精确度高。 **说明** 只在**使用场景**为**文本反垃圾**时支持。

参数	说明
匹配方式	**文本类型**为**关键词**时，选择文本库的匹配方式。取值： ○ **精确匹配**：待检测文本中包括与库中的词完全一样的内容时才命中。 ○ **模糊匹配**：待检测文本以及关键词都会经过预处理，预处理后进行匹配。预处理的逻辑如下： ■ 字母大写统一转换为小写。例如，输入检测文本“bitCoin”，会命中关键词“bitcoin”。 ■ 繁体中文统一转换为简体。例如，输入检测文本“中國”，会命中关键词“中国”。 ■ 相似字转换。例如，输入检测文本“②”，会命中关键词“2”。 **说明** 相似文本类型的文本库默认使用模糊匹配。
识别结果	选择文本库的用途。 ○ **文本类型**为**关键词**时，取值： ■ **黑名单**：若命中文本库中样本，则机审结果返回违规（suggestion=block）。 ■ **疑似名单**：若命中文本库中样本，则机审结果返回疑似（suggestion=review）。 ■ **忽略名单**：不检测文本库中样本，但是会检测除了样本库以外的其他内容。 ○ **文本类型**为**相似文本**时，取值： ■ **黑名单**：若命中文本库中样本，则机审结果返回违规（suggestion=block）。 ■ **疑似名单**：若命中文本库中样本，则机审结果返回疑似（suggestion=review）。 ■ **白名单**：若命中文本库中样本，则机审结果返回正常（suggestion=pass）。
BizType	BizType允许您根据不同的业务需求配置并应用不同的文本库，例如使用BizType指定在某次检测中应用文本库。BizType生效逻辑如下： ○ 文本库设置BizType为“A”，且API检测请求中传递了BizType为“A”，则检测文本只会使用BizType为“A”的文本库（前提是文本库已开启）。 ○ 其他情况下，检测文本均会使用所有已开启的文本库。

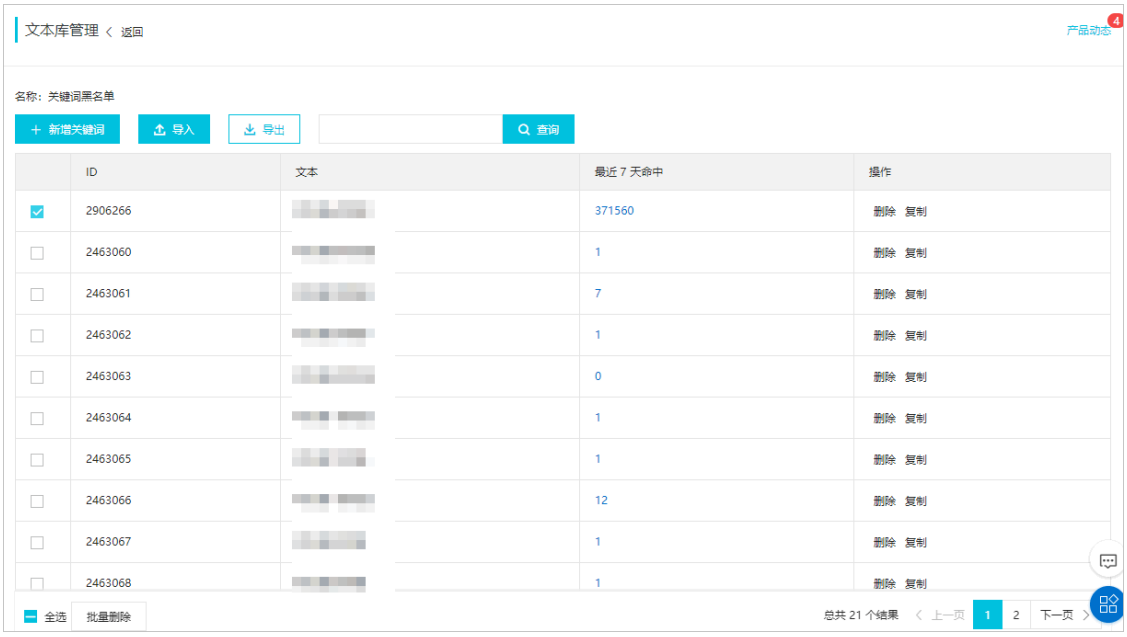
成功创建文本库后，您可以在文本库列表中查看新建的文本库。

5. 管理文本库中的文本（关键词、相似文本）。

   自定义文本库列表显示所有自定义文本库，其中系统回流文本库的名称前有**系统**标识，且按照“使用场景_FEEDBACK_WHITE/BLACK”命名。例如，**ANTISPAM_FEEDBACK_BLACK**是由系统回流生成的用于文本反垃圾场景的黑名单文本库。

   i. 定位到要操作的关键词文本库，单击其操作列下的**管理**。

ii. 在**文本库管理**页面，维护文本库内的关键词。



文本库管理页面展示了所有已添加的关键词信息和**最近7天命中**次数（不包括当天的命中数据）。

> **说明** 在文本库新增、删除关键词，大约需要15分钟生效。

- 单击**新增关键词**或**导入**，按照页面提示在文本库中增加关键词。

- 选中不需要的关键词，单击**批量删除**，删除关键词。您也可以单击不需要的关键词下的**删除**，单独将其删除。

相关操作

对于自定义文本库（非系统回流文本库），您可通过**自定义文本库**页面中**操作**列下的**删除**、**修改**和**停用**对目标文本库进行操作。

2.5. 自助审核

内容安全控制台中呈现了内容检测API检测出的数据结果。针对您的业务场景，您可以对机器的检测结果进行二次人工审核。人工审核后，下次同样的检测内容识别出的结果会与您设置的结果保持一致。本文介绍了使用自助审核的具体操作。

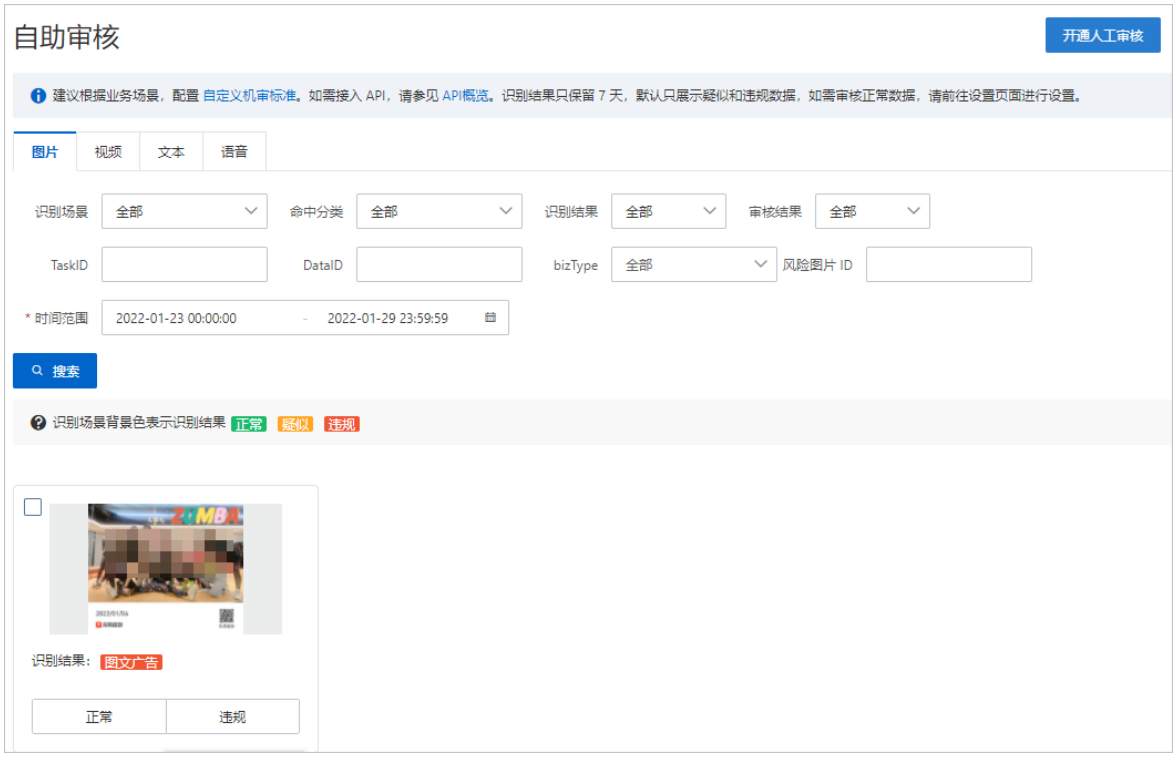
背景信息

- 自助审核默认只展示机器审核结果为疑似（review）或者违规（block）的数据。如需展示机审结果正常（pass）的数据，请在控制台上进行设置，具体请参见自定义机审标准。
- 图像、视频、语音、文本均可以进行人工审核，但只有图像、文本的自助审核结果会自动回流入风险样本库。
- 机器的检测数据只保留最多7天，请及时处理。

操作步骤

1. 登录内容安全控制台。
2. 在左侧导航栏，选择**机器审核 > 自助审核**页面。
3. 通过页签选择要审核的内容的类型（**图片**、**视频**、**文本**、**语音**），单击进入审核页面。
4. 按照以下方式进行标记。

> **说明** 仅以图片标记为例，其他类型内容的标记方法与之类似。



- 对于您认为正常，却被识别为违规（block）或者疑似（review）的样本，将其标记为**正常**。
- 对于您认为需要管控，却被识别为疑似（review）或者正常（pass）的样本，将其标记为**违规**，并选择违规原因：**涉黄**、**暴恐涉政**、**图文广告**、**不良场景**。

- 支持勾选多张图片后进行批量处理，如**批量标记正常**、**批量标记违规**。
- 单击样本图片，查看其**详细信息**。

被标记样本以及类似样本的检测结果将会按照您的标记结果实时纠正，且自动回流入对应的样本库中。

如下图所示带有红色**系统**标识的图库即为系统回流图库（文本库与之类似）。

2.6. 检测结果

您可以在检测结果页面查看机器审核的结果。

操作步骤

1. 登录内容安全控制台。
2. 在左侧导航栏，选择**机器审核 > 检测结果**。

3. 在**检测结果**页面，设置查询条件查询检测结果。

   您可以查看**图片**、**视频**、**文本**、**语音**的检测结果。

   支持的查询条件包含：**识别场景**、**命中分类**、**识别结果**、**时间范围**、TaskID、DataID、bizType、**风险图片ID**。

2.7. 数据统计

您可以在内容安全控制台查看内容检测API的调用统计数据。

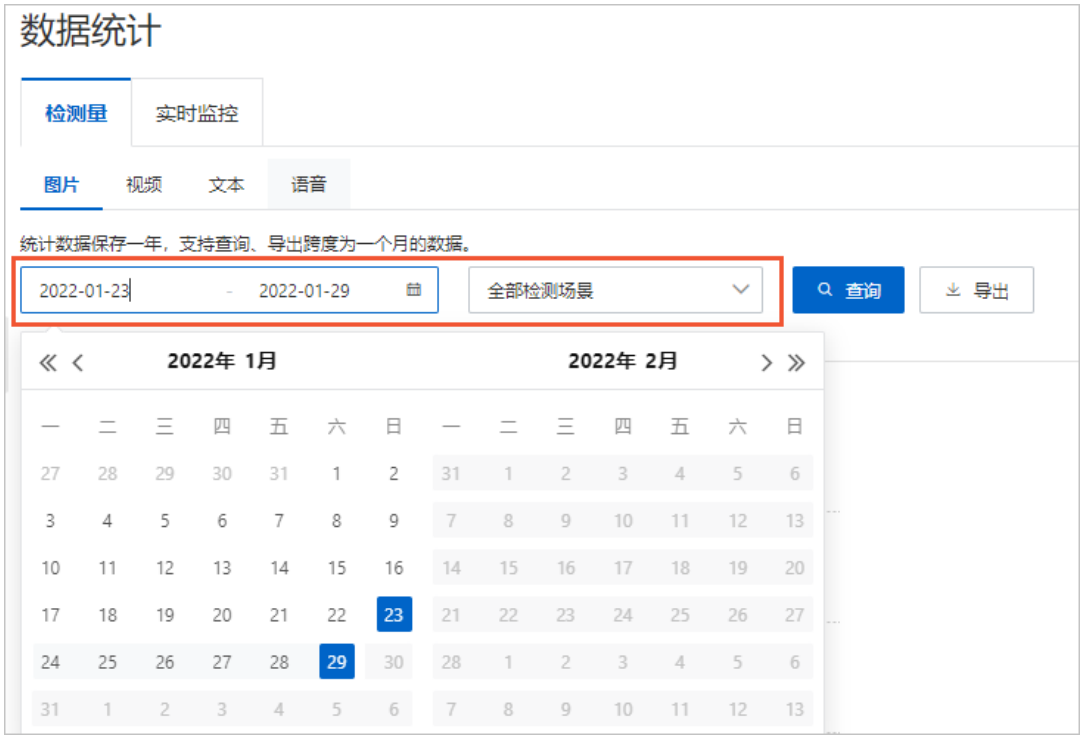
背景信息

内容安全控制台汇总了内容检测API的调用统计数据，支持查询最近1年内图片、视频、文本、语音检测接口的总调用次数，以及不同检测场景下检测结果（确认违规量、疑似违规量、正常量）的分布信息。

操作步骤

1. 登录内容安全控制台。
2. 在左侧导航栏，选择**机器审核 > 数据统计**。
3. 通过页签选择要查询的检测接口类型：**图片**、**视频**、**文本**、**语音**。
4. 在**数据统计**页面，选择查询时间，并单击**查询**。

   支持查询的时间段为最近1年内。支持设置的时间跨度为1个月。

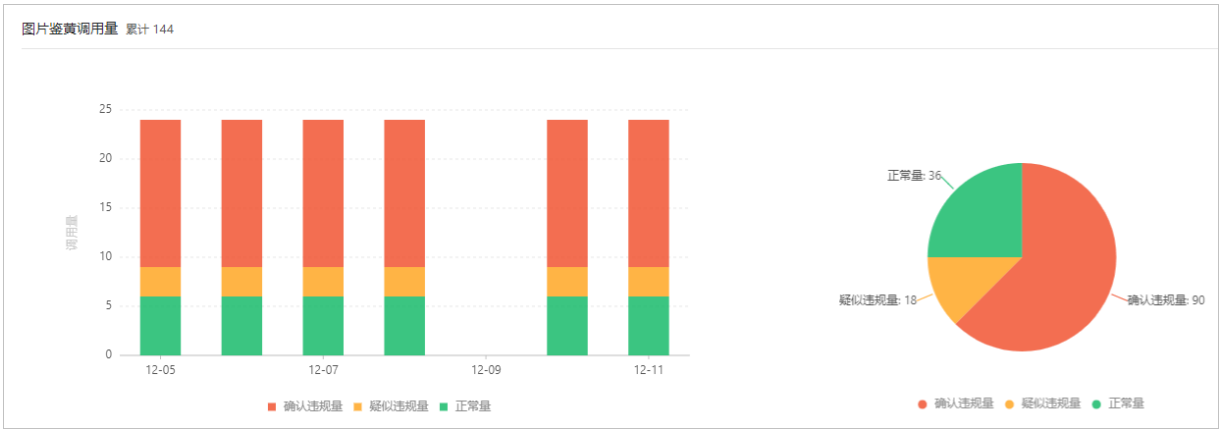


5. 分检测场景查看调用统计报表，并根据需要导出报表。
   - 报表说明
   - 导出说明

报表说明

以**图片鉴黄调用量**报表为例，报表展示了每日调用图片鉴黄检测接口（scene=porn）的次数，并统计了不同检测结果的数量和分布。图例说明如下：

- **确认违规量**：返回suggestion=block的检测请求的数量。
- **疑似违规量**：返回suggestion=review的检测请求的数量。
- **正常量**：返回suggestion=pass的检测请求的数量。



不同检测对象的调用量单位不同，具体说明如下：

- 图片检测：调用量单位是图片的数量（张）。
- 视频检测：调用量单位有两种，一种是视频的截帧数（张），一种是视频的时长（分钟）。
- 文本检测：调用量单位是文本行数（条）。
- 语音检测：调用量单位是语音的时长（分钟）。

导出说明

导出的报表是Excel格式。导出数据的时间范围与您设置的查询条件一致。导出的Excel只包含有调用量的检测场景（对应API接口调用时传递的scene参数值），每个场景对应一张表单，表单中按天记录调用量。

表单中出现的行头字段说明见下表。

名称	含义	单位
day	调用日期	无
totalImageCount	检测图片总量	张
blockImageCount	违规图片量	张
reviewImageCount	疑似违规图片量	张
passImageCount	正常图片量	张
totalVideoCount	检测视频总量	个
blockVideoCount	违规视频量	个
reviewVideoCount	疑似违规视频量	个
passVideoCount	正常视频量	个
innerFrameCount	视频的系统截帧总量	张
outerFrameCount	视频的用户截帧总量	张
totalTextCount	检测文本总量	条
blockTextCount	违规文本量	条
reviewTextCount	疑似违规文本量	条
passTextCount	正常文本量	条
totalVoiceDuration	检测语音总量	分钟
blockVoiceDuration	违规语音量	分钟
reviewVoiceDuration	疑似违规语音量	分钟
passVoiceDuration	正常语音量	分钟

2.8. 样本反馈

在使用内容检测API过程中，如果您发现内容安全的算法结果在您的业务中被认为是不符合预期的，您可以通过反馈接口将样本回流给我们。

如果您有自己的审核平台，您可以直接对接反馈接口，将审核后认为识别有误的样本回流给我们。在收到您的反馈后，我们会在下个版本的模型迭代中将您的反馈数据加入训练。训练后的模型对您的业务场景更具有适应性。

样本反馈目前支持图片样本反馈、视频样本反馈和文本样本反馈，具体使用方式请参见相关文档。

自动加入自定义图库

模型训练需要积累足够的样本，因此可能无法立即生效。如果有需要，您可以开启自动加入自定义图库功能，实时纠正结果。

针对图片样本的反馈，您可以在内容安全控制台直接管理回流图像库。操作方法类似其他自定义图库，但是不支持创建与删除回流图像库。关于自定义图库的操作方法，请参见自定义图库。

- 对于您认为正常的样本，在反馈接口的label字段中传入*normal*，可以将该样本加入白名单。
- 对于您认为违规的样本，在反馈接口的label字段中传入任意字段（建议您使用*porn*、*ad*、*terrorism*等风险字段），可以将该样本加入黑名单。

2.9. 授权访问MTS服务

内容安全需要通过调用媒体处理（ApsaraVideo for Media Processing，原MTS）才能对上传到OSS的视频进行视频截帧，所以您必须先授权MTS服务以内容安全的身份递交视频截帧任务。该操作通过阿里云访问控制中的角色管理功能实现，本文介绍了您需要完成的步骤。

背景信息

- 在提交视频内容异步检测任务时，如果您选择通过OSS地址上传视频URL的方式，则内容安全对上传的OSS视频自动截帧。您必须完成授权操作，才可以保证自动截帧正常。
- 目前OSS支持的区域（Region）包括：华东1（杭州）、华东2（上海）、华北2（北京）和华南1（深圳）。

操作步骤

1. 创建RAM角色并授权。

   在您的阿里云账号下创建MTS服务角色，并指定由开通内容安全的阿里云账号扮演该角色。

   i. 登录RAM控制台。
   ii. 前往**RAM角色管理**页面，单击**创建RAM角色**。
   iii. 选择可信实体类型为**阿里云服务**，并单击**下一步**。
   iv. 配置**角色名称**，并选择受信服务为**多媒体转码服务**。
   v. 单击**完成**。

   > **说明** 该操作可能需要通过手机验证。

   vi. 成功创建角色后，单击**为角色授权**。
   vii. 在**添加权限**页面，为角色授予**AliyunOSSReadOnlyAccess**系统权限策略，并单击**确定**。
   该操作授权服务角色以只读权限访问您的阿里云账号下的OSS数据。

2. 修改服务角色的信任策略。

   i. 在**RAM角色管理**页面，单击新建的角色名称进入角色详情。
   ii. 打开**信任策略管理**页签，并单击**修改信任策略**。
   iii. 在**修改信任策略**页面，将 `"Service"` 下的内容修改为 `"118484706224****@mts.aliyuncs.com"` ，并单击**确定**。
   该操作指定由内容安全的阿里云账号（UID：118484706224****）扮演所创建的服务角色，调用其MTS服务。

后续步骤

完成授权操作后，您提交的OSS视频异步检测任务时，需要上传规定格式的URL作为检测对象，具体操作如下：

1. 在角色详情中，查看并复制角色的**ARN**（Aliyun Resource Name，阿里云全局资源名称）。
2. 对要检测的OSS视频对象，按照以下格式拼接生成视频URL： `oss://arn@bucket.region/object`

例如，假设您在深圳OSS的Bucket *f***上有视频对象*video/bar.mp4*需要检测，则拼接生成的URL为 `oss://acs:ram::1184847062224****:role/mts-to-a@f**.cn-shenzhen/video/bar.mp4` （ `1184847062224****` 是您的16位阿里云账号ID。）

3. 提交视频异步检测任务时，上传拼接生成的URL作为检测对象。

> **说明**
>
> - 仅在使用视频异步检测时支持上传视频URL的方式。更多信息，请参见异步检测。
> - 目前支持的区域（Region）包括：华东1（杭州）、华东2（上海）、华北2（北京）和华南1（深圳）。

3.人工审核

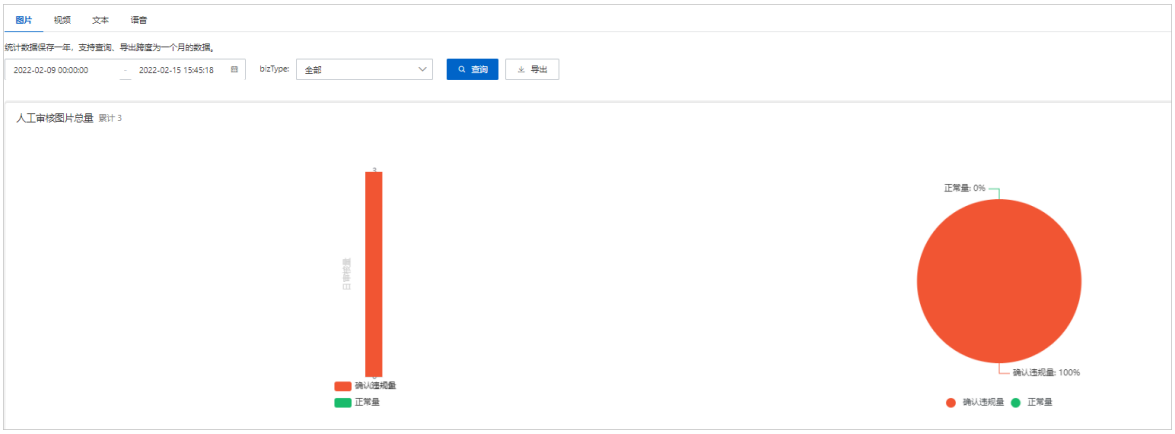
3.1. 查看数据统计报表

您可以在内容安全控制台查看人工审核的统计数据。

操作步骤

1. 登录内容安全控制台。
2. 在左侧导航栏，选择**人工审核 > 数据统计**。
3. 在**数据统计**页面，通过页签选择要查询的检测接口类型：**图片**、**视频**、**文本**、**语音**。
4. 选择查询的时间范围、**检测场景**，并单击**查询**。

   支持查询的时间段为最近1年内。支持设置的最大时间跨度为1个月。



5. 分检测场景查看调用统计报表。

   图例说明如下：

   - **确认违规量**：返回suggestion=block的检测请求的数量。
   - **正常量**：返回suggestion=pass的检测请求的数量。

   不同检测对象的调用量单位不同，具体说明如下：

   - 图片检测：调用量单位是图片的数量（张）。
   - 视频检测：调用量单位有两种，一种是视频的截帧数（张），一种是视频的时长（分钟）。
   - 文本检测：调用量单位是文本行数（条）。
   - 语音检测：调用量单位是语音的时长（分钟）。

   您也可以通过导出报表查看相关内容，导出的报表是Excel格式。导出数据的时间范围与您设置的查询条件一致。导出的Excel只包含有调用量的检测场景（对应API接口调用时传递的scene参数值），每个场景对应一张表单，表单中按天记录调用量。

名称	含义
day	调用日期
blockCount	确认违规量
passCount	正常量
totalCount	总量

名称	含义
totalDuration	语音时长

3.2. 接入人工审核服务

内容安全提供7*24小时的人工审核服务（人机审核和纯人工审核）。人机审核服务可返回机器识别结果，并根据机器识别的结果选择性进行人工审核。纯人工审核服务对全量的内容数据进行人工审核，仅返回人工审核的结果，不返回机器识别结果。

前提条件

- 已开通内容安全服务。如果您未开通内容安全服务，则无法执行本文操作。

  > **说明** 人工审核服务为收费服务，具体计费方式，请参见内容安全产品定价。

- 已完成内容检测API开发准备工作，请参见内容检测API开发准备。
- 已联系阿里云技术支持人员对人工审核的数据量、时间范围和审核标准进行沟通。您可以申请人工审核保障服务。

（推荐）人机审核操作步骤

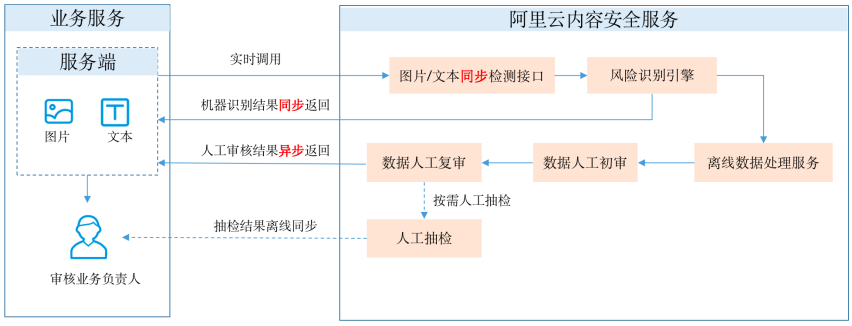
1. 登录内容安全控制台。
2. 在**设置 > 内容检测API**页面，配置审核结果回调地址。

   具体操作请参见配置消息通知。
3. 根据您的业务场景，调用指定的内容安全服务接口。

   具体说明如下：

   - 图片审核场景

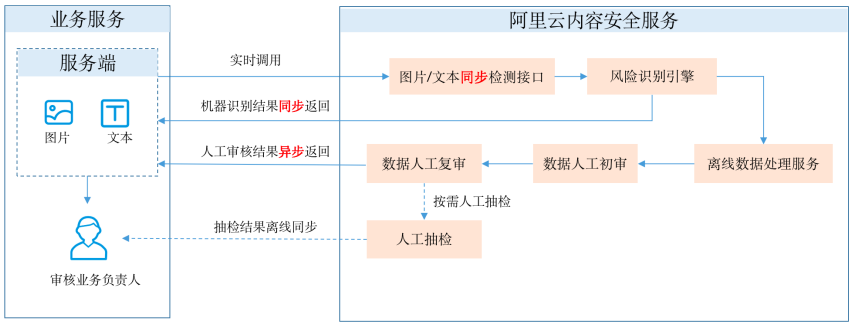
     您需要调用图片同步检测接口，业务流程如下图所示。



     图片同步检测接口的识别结果会同步返回到业务服务端，人工审核结果会通过回调的方式返回到您配置的审核结果回调地址。

   - 文本审核场景

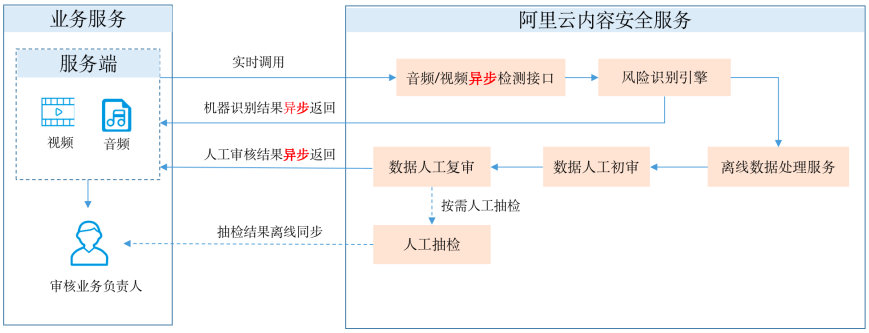
     您需要调用文本同步检测接口，业务流程如下图所示。



文本同步检测接口的识别结果会同步返回到业务服务端，人工审核结果会通过回调的方式返回到您配置的审核结果回调地址。

- 视频审核场景

  您需要调用视频异步检测接口，业务流程如下图所示。

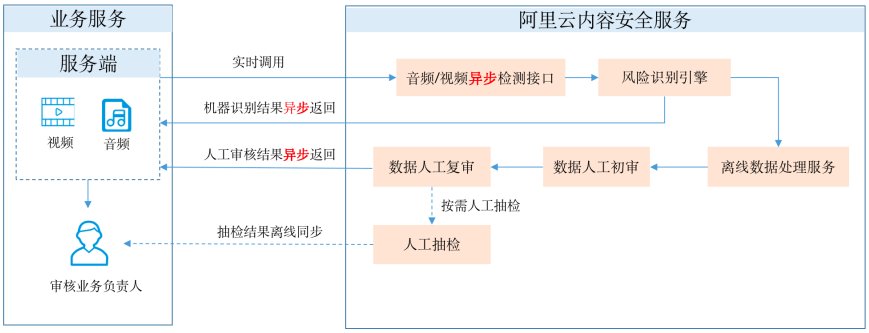


视频异步检测接口的识别结果会通过异步消息的方式返回到业务服务端（也支持您主动调用结果查询接口获取结果），人工审核结果会通过回调的方式返回到您配置的审核结果回调地址。

人工审核结果会在视频异步检测接口识别完成后返回，如果您在获取结果前调用了停止视频检测接口停止视频异步检测，则人工审核结果不会返回。

- 音频审核场景

  您需要调用音频异步检测接口，业务流程如下图所示。



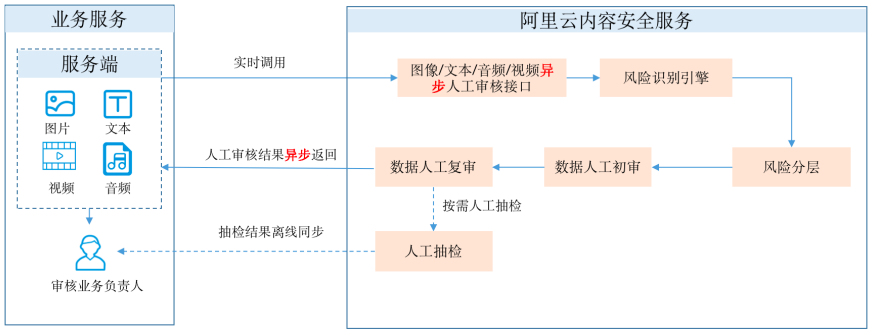
音频异步检测接口的识别结果会通过异步消息的方式返回到业务服务端（也支持您主动调用结果查询接口获取结果），人工审核结果会通过回调的方式返回到您配置的审核结果回调地址。

人工审核结果会在音频异步检测接口识别完成后返回，如果您在获取结果前调用了取消音频检测接口取消音频异步检测，则人工审核结果不会返回。

纯人工审核操作步骤

1. 根据文件类型（图片、视频、音频和文本）选择对应的人工审核接口，在接口中设置接收人工审核结果的回调地址（callback信息）。

2. 提交内容进行人工审核，等待人工审核结果返回再进行业务处理，业务流程如下图所示。



人工审核的接口，请参见图片人工审核和文本人工审核。

3.3. 复核人工审核结果

如果内容安全的人工审核结果不符合您的预想，您可以对审核结果进行复核。文本介绍如何复核人工审核的结果。

背景信息

人工审核模块展示所有人工审核的数据，从审核模式来分，包括人机审核和纯人工审核。从审核结果来分，包括**正常**（pass）、**疑似**（review）和**违规**（block）的数据。

操作步骤

1. 登录内容安全控制台。
2. 在左侧导航栏，选择**人工审核 > 人工审核&复核**。
3. 单击待审核的内容类型（包含**图片**、**视频**、**文本**、**语音**），按照如下方式复核人工审核结果。

功能区	功能描述	支持的操作
条件查询区（图示①）	根据您的业务需要，设置查询条件，查找待复核的内容。 目前支持按照**流入时间**、**违规类型**、**人审结果**、**业务场景**、**复核结果**、**复核违规类型**、**TaskID**、**DataID**。	设置好查询条件后，单击**搜索**。

功能区	功能描述	支持的操作
结果展示区域（图示②）	根据查询条件查询到的人工审核结果显示在该区域，您可以根据实际业务，对人工审核的结果进行复核。	◦ 对于您认为内容是正常的，但是被阿里云人工审核为违规的数据，可以手动将其标记为正常。 ◦ 对于您认为内容需要被管控，但是被阿里云人工审核为正常的数据，可以手动将其标记为违规，并选择违规原因（包含**涉政**、**涉黄**、**暴恐**、**图文违规**）。 支持对内容执行单条复核和批量复核两种方式： ◦ **单条复核**：勾选1条数据（图片、视频、文本、语音），在**复核**列进行复核，如标记为**正常**、**违规**。 ◦ **批量复核**：勾选多条数据（图片、视频、文本、语音），在页面最下方进行批量复核，如**批量标记正常**、**批量标记违规**。 单击目标数据，展开**详细信息**面板，查看对应数据的详细内容。

当您通过控制台复核结果并确认后，系统会自动对复核结果进行回调通知。

3.4. 配置回调通知

内容安全支持以异步消息通知的方式向您发送人工审核和人工审核复核的结果。如果您的业务需要使用或集成相关数据，则可以配置回调通知。本文介绍了配置人工审核及复核人工审核回调通知的操作方法。

背景信息

人工审核回调分为阿里云人工审核回调通知和复核人工审核回调通知：

- 人工审核回调通知：内容安全人工审核团队完成人工审核后，服务端将识别的结果以POST请求的方式，发送到您设置的HTTP回调通知地址。
- 复核人工审核

回调通知：内容安全客户通过控制台完成复核确认后，服务端将标记的结果以POST请求的方式，发送到您设置的HTTP回调通知地址。

相关概念

在配置回调通知前，请了解下表中描述的相关概念。

名称	说明
回调地址	回调地址是您在内容安全控制台配置的服务端地址，通常是您自己的业务服务器的公网地址。回调地址需要满足以下要求： • 应为HTTP、HTTPS协议接口的公网可访问的URL。 • 支持POST方法。 • 支持传输数据编码采用UTF-8。 • 支持数据接收格式为 `application` 、 `x-www-form-urlencoded` 。 • 支持表单参数checksum和content。
seed	seed值用于校验发送到您设置的HTTP回调通知地址的请求是否来自内容安全服务端。
回调次数	您的服务端接收到内容安全推送的回调消息后，如果返回的HTTP状态码为200，表示接收成功；如果返回其他的HTTP状态码，均视为接收失败。如果接收失败，内容安全服务端会重复推送回调消息，直至您接收成功。内容安全最多重复推送16次回调消息。

名称	说明
回调数据	回调数据是内容安全服务端向您设置的回调通知地址返回的数据内容。关于回调数据的结构描述，请参见回调通知表单数据。

回调通知表单数据

名称	类型	描述
checksum	String	总和校验码，由 <阿里云账号ID> + <seed> + <content> 拼成字符串，通过SHA256算法生成。其中，阿里云账号ID可以在阿里云控制台上查询。 **说明** 为防篡改，您可以在获取到推送结果时，按上述算法生成字符串，与checksum做一次校验。
content	String	字符串格式保存的JSON对象，请自行解析反转成JSON对象。关于content解析成JSON后的结构，请参见下文content表单字段说明。 **说明** 在内容检测API和OSS违规检测中，content的参数结构不同。

人工审核回调通知

人工审核接口不支持实时返回检测结果，系统支持回调方式通知调用方。人工审核分为人机审核和纯人工审核，两者的回调方式有一定的差异，以下分别进行介绍。

- **人机审核**

  如果您使用的是阿里云人工审核服务+机审服务，人工审核模块的回调需要通过控制台配置消息通知来实现，具体操作步骤如下：

  i. 登录内容安全控制台。
  ii. 在左侧导航栏，选择**设置 > 机器审核**。
  iii. 在**机器审核**页面，单击**消息通知**。
  iv. 在**消息通知**页签，单击**新增通知方案**。
  v. 在**新增通知方案**对话框，输入**方案名称**、**回调地址**，选择**加密算法**、**通知类型**、**审核结果**。然后单击**确认**。

  保存完成后，系统自动生成seed。seed值用于校验您的回调接口收到的请求来自阿里云。请保存自动生成的seed。

> 注意
>
> - 如果您已经配置了机审回调通知服务，您可以直接复用之前的配置，也可以根据业务需要重新配置。
> - 系统支持对于所有的审核结果进行回调，您也可以根据自己的业务场景选择部分审核结果进行回调。
> - 通知类型项请选择**阿里云人审结果**。
> - 加密算法方式：
>   - **SHA256**：使用HMAC-SHA256加密算法。
>   - **国密SM3**：使用国密SM3加密算法，返回十六进制的字符串，且字符串由小写字母和数字组成。
>
>   例如，abc经国密SM3加密后返回66c7f0f462eeedd9d1f2d46bdc10e4e24167c4875cf2f7a2297da02b8f4ba8e0。

vi. 在**机器审核**页面，单击**业务场景管理**。

vii. 在**业务场景管理**页签，单击目标业务场景右侧**操作列**的**关联消息通知**，关联您配置的**回调通知方案**。

- **纯人工审核**

  如果您使用的是阿里云纯人工审核服务，人工审核模块的回调需要通过请求参数传入callback（回调地址）和seed（校验字符串）来实现。

  i. 自行准备好接收扫描结果的HTTP回调地址和seed参数。

  ii. 调用内容检测异步API接口时，传递相应的callback和seed请求参数。

  关于具体参数说明，请参见图片人工审核、视频人工审核、文本人工审核、复核人工审核结果。

content表单字段说明

启用回调通知后，内容安全将按照回调配置发送内容检测API回调通知。回调通知中包含content表单数据。下表描述了content表单字段的结构。

content表单字段结构描述

名称	类型	是否必须	说明
humanAuditResult	JSONObject	否	阿里云的人工审核结果。如果您购买了阿里云的人工审核服务，则人工审核的结果体现在该字段中。具体请参见humanAuditResult。

humanAuditResult

名称	类型	是否必须	说明
suggestion	String	是	阿里云人工审核的结果，取值： • **block**：阿里云人工审核结果为违规。 • **pass**：阿里云人工审核结果为正常。
customAuditFlag	Boolean	是	• **false**：阿里云人工审核结果。 • **true**：客户控制台人工审核复核结果。 **说明** 此标志位用来区分返回报文内容是阿里人工审核结果或者客户控制台人工审核复核结果。

名称	类型	是否必须	说明
taskId	String	是	检测任务的ID。通过任务ID可以关联到对应内容的审核结果。
dataId	String	是	检测内容的ID。
labels	StringArray	否	人工审核的标签结果，可能有多个值。 **说明** 默认不返回该参数。配置该字段需要额外收费，具体请联系商务沟通。

content示例

```
{
    "humanAuditResult": {
          "suggestion": "pass",
          "customAuditFlag": false,
          "dataId": "yyyy",
          "labels": [
              "色情",
              "低俗"
        ],
          "taskId": "xxxxxx"
    }
}
```

4.站点检测

4.1. 使用简介

站点检测服务帮助您定期检查网站首页和全站内容，及时发现您的网站在内容安全方面可能存在的风险（例如，首页篡改、挂马暗链、色情低俗、涉政暴恐等），并向您展示违规内容的具体地址，帮助您查看和修复。

功能描述

站点检测的对象是您的网站上的网页和图片，以URL数量进行计数。在单个网站的一个检测周期内，站点检测支持的最大检测容量为10万个URL。站点检测包含首页检测和全站检测功能。

- 首页检测：定期对您网站的首页进行检测，展示最近一次的检查结果。检查结果涵盖首页篡改、挂马暗链、色情低俗、涉政暴恐等风险提示，并提供源码、文本、图片三类呈现方式，供您参照和整改。
- 全站检测：定期对您网站域名下的网页进行自动化全站内容检测，展示最近一次的检查结果。检查结果涵盖挂马暗链、色情低俗、涉政暴恐等风险提示，并提供源码、文本、图片三类呈现方式，供您参照和整改。

使用流程

站点检测服务定期检查您的网站首页和全站内容，及时发现您的网站在内容安全方面可能存在的风险（例如首页篡改、挂马暗链、色情低俗、涉政暴恐等），并向您展示违规内容的具体地址，帮助您查看和修复。您可以设置消息通知，获取实时的站点首页风险提醒。

在使用站点检测功能之前，您需要先购买站点检测实例。购买实例后，需要将实例绑定到您的站点、添加要检测的网站域名和首页地址、设定首页和全站检测的频率，并完成网站鉴权。完成设置后，系统将定期按照您设定的频率对首页和全站内容（包含网页源码、文本和图片）进行检测。如果发现有风险，将按照您设定的消息接收方式通知您，您也可以登录内容安全控制台查看检测结果。

站点检测功能只支持包年包月的计费方式。具体价格信息，请参见站点检测购买页。

序列	操作	详细说明
步骤一	创建站点检测任务	在使用站点检测功能之前，您需要购买站点检测实例，并将实例绑定到待检测的站点上。 关于创建站点检测任务的具体操作，请参见创建站点检测任务。
步骤二	配置消息通知	在使用站点检测服务时，您可以设置风险通知方式，设置风险通知后，系统会按照设置条件把检测结果发送给您。 关于配置消息通知的具体操作，请参见配置消息通知。
步骤三	配置风险库	在使用站点检测服务时，您可以添加自定义关键词或者自定义图片进行黑名单、白名单防控。 关于配置风险库的具体操作，请参见配置风险库。
步骤四	查看检测结果	系统将定期按照您设定的频率对首页和全站内容（包含网页源码、文本和图片）进行检测。如果发现有风险，将按照您设定的消息接收方式通知您。您也可以登录内容安全控制台查看检测结果。 关于查看检测结果的具体操作，请参见查看检测结果。

4.2. 创建站点检测任务

在使用内容安全站点检测功能前，您需要先创建一个站点检测任务，用来设置您需要检测的站点信息、检测频率，并进行网站鉴权。设置完成后，内容安全站点检测功能会定期对您的站点进行检测。

操作步骤

1. 购买站点检测实例。

   每个实例只能绑定一个站点（建议您填写站点的根域名），如果需要绑定多个站点，需要您购买对应数量的实例。

   i. 登录内容安全控制台。

   ii. 在左侧导航栏，选择**设置 > 站点检测**。

   iii. 在**实例管理**页签，单击**购买实例**。

   iv. 在**站点检测（包年）**页面，根据实际需求配置相关参数，单击**立即购买**完成订单。

2. 对实例绑定需要检测的站点。

   i. 在**实例管理**页签，选择**未绑定**状态的实例，单击右侧操作列的**绑定站点**。

   ii. 在**绑定站点**对话框中，请参考下表各配置项的介绍配置**协议**、**域名**、**默认首页地址**、**首页检测间隔**和**全站检测频率**，并单击**下一步**。绑定站点页面说明表

配置项	说明
协议	选择协议类型，取值： ■ **HTTP**：超文本传输协议。 ■ **HTTPS**：由HTTP加上TLS或SSL协议构建的可进行加密传输、身份认证的网络协议。 **说明** 如果您的网站通过HTTP和HTTPS协议分别响应不同的内容，而且内容差异较大，建议您使用两个实例，分别绑定网站的HTTP协议和HTTPS协议。
域名	您站点的域名。在域名中不要包含 `http://` 或者 `https://` 。如果您的网站有多个子域名，建议您填写根域名。
默认首页地址	站点首页的完整网址，输入的网址必须在您要绑定的域名下。
首页检测间隔	首页检测时间间隔，每隔多少小时对您的网站首页进行一次检测。
全站检测频率	执行全站检测的频率，取值： ■ **低：7天1次** ■ **高：1天1次** **说明** 站点检测频率越高，检测占用的带宽及产生的带宽费用也越多。如果您的网站内容比较多，且网站带宽不足的话，过高的检测频率可能影响您网站的正常访问速度。如果您不希望影响网站性能，建议您选择较低的检测频率。

3. 选择站点验证方式。

   在**验证站点**对话框，请参考下表各验证方式的介绍选择一种验证方式，然后单击**立即验证**。

验证方式	说明
阿里云账户验证	验证待检测站点（域名）是否在您当前登录的阿里云账号的资产下。
主机文件验证	根据页面提示在域名对应主机的根目录下生成相应的文件进行验证。
CNAME验证	根据页面提示在待检测域名的解析记录中增加指定的CNAME记录进行验证。
网站首页HTML标签验证	根据页面提示修改网站首页HTML源文件进行验证。

站点验证主要是为了防止出现未经授权的检测场景。如果您暂时不方便进行验证，您可以单击**稍后验证**，保存当前已输入的数据。验证通过后，完成站点绑定和检测设置，目标实例自动开始检测。

4. （可选）在**实例管理**页面，单击状态为**检测中**的实例，在右侧**操作**列下，根据需要执行以下操作。

操作项	说明
暂停/启动检测	如果您不希望在当前时间执行检测，您可以暂停检测。已暂停的检测，通过启动检测继续进行之前的检测任务。
设定首页防篡改基准	确认当前首页基准。如果您想更换首页基准，可以通过**重新获取当前首页**重新抓取当前首页作为首页基准。 **说明** 绑定站点开启检测时，站点检测任务会抓取当前首页作为判断首页是否被篡改的基准。若您更新过首页内容，建议您设置首页防篡改基准，方便站点检测任务重新抓取当前首页。
添加重点监控URL	输入您待检测的URL，每行一个URL，使用回车换行。最多支持添加5000个URL。如果您的网站内容很多，您担心在检测中重要的URL会被遗漏，您可以自定义重点监控URL，站点检测任务会优先检测您添加的URL。
编辑站点	修改实例绑定的站点和检测频率信息。
重新验证	如果您的验证失效或在步骤3中选择**稍后验证**，您可以重新验证对站点的管理权。
续费	为实例续费，可以延长其使用时长。
解除绑定	如果您不希望继续向已绑定的站点提供检测服务，您可以解除绑定。 **说明** 解除绑定后，已购买的实例不会释放，但是您可以将其绑定到别的站点，为别的站点提供检测服务。

后续步骤

查看检测结果。

4.3. 配置消息通知

内容安全支持消息通知，当检测出您的站点存在风险时，会根据您的设置通知提醒您，默认每天推送一次消息。您可以设置消息接收方式、账号和接收时间，也可以开启或关闭首页风险实时通知。

操作步骤

1. 登录内容安全控制台。
2. 在左侧导航栏，选择**设置 > 站点检测**。
3. 单击**消息通知**。
4. 在**消息通知**页签，填写风险预警的**通知账户**信息，选择相应的**提醒方式**，并设置**所在时区**、**推送时间**和**实时消息提醒**。

> **说明** 开启**实时消息提醒**后，系统一旦检测出首页存在风险，会实时发送消息给您。多次检测到风险时，为避免您被打扰，针对单个域名每天最多发送一次提醒。

5. 单击**保存**，完成消息通知配置。

4.4. 配置风险库

在站点检测时，如果您需要对特殊的词汇进行专门识别和防控，您可以添加自定义关键词进行黑名单防控，进行黑名单防控。在使用站点检测检查图片时，您可以将特定图片定义为白名单、黑名单图片，进行过滤、防控。

背景信息

在使用站点检测服务时，关键词只支持UTF-8格式，且添加的关键词或者图片会在15分钟内生效。例如，您在添加自定义文本后，该文本内容会在15分钟内应用到您的检测任务中。

配置自定义文本库

1. 登录内容安全控制台。
2. 在左侧导航栏，选择**站点检测 > 风险库管理**。
3. 在**自定义文本库**页签，单击**创建文本库**。

   最多支持创建10个文本库。
4. 在**创建自定义文本库**对话框中，设置**名称**，选择要应用该文本库的实例，然后单击**确定**。

5. 在**自定义文本库**页签，选择新创建的文本库，在**操作**选项中，单击**管理**。

6. 在**文本库管理**页面，单击**新增关键词**。
7. 在右侧的**新增关键词**面板，按照页面提示输入或导入关键词，并单击**确定**，完成添加关键词。

配置自定义图库

1. 登录内容安全控制台。
2. 在左侧导航栏，选择**站点检测 > 风险库管理**。
3. 在**自定义图库**页签，单击**创建图库**。

   最多支持创建10个图库。
4. 在**创建图库**对话框中，请参考下表配置项的介绍设置图库的**名称**、**使用场景**、**识别结果**和**实例**，并单击**确定**。

创建图片页面说明表

配置项	说明
名称	用于识别此图库的名称。

配置项	说明
使用场景	支持**鉴黄**和**暴恐**场景。
识别结果	支持**黑名单**和**白名单**识别结果。黑名单图库用于特殊防控不良图片，白名单图库会在检测中忽略并过滤您添加的图片。
实例	应用该图库的实例。 **说明** 在检测站点时，只有您在此处选择的实例才会使用该图库进行站点检测。

5. 在**自定义图库**页签，选择新创建的图库，在**操作**选项中，单击**管理**。
6. 单击**选择文件**，并上传本地图片至当前图库。

> **说明** 每个图库支持添加最多10000张图片。

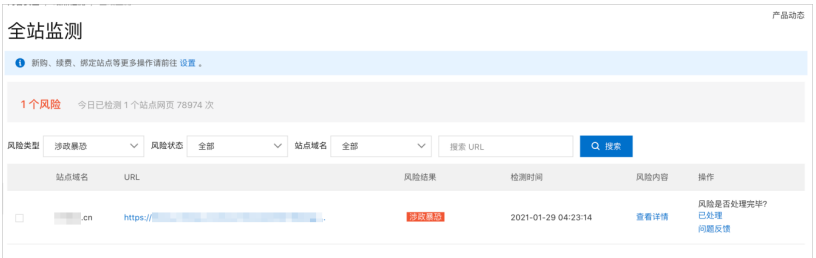
4.5. 查看检测结果

内容安全站点检测功能将定期按照您设定的频率对首页和全站内容（包含网页源码、文本和图片）进行检测。您可以在内容安全控制台查看首页和全站检测结果。

操作步骤

1. 登录内容安全控制台。
2. 在左侧导航栏，选择**站点检测 > 首页监测**或者**站点检测 > 全站监测**。

   查看最近一次检测结果。



3. 定位到存在风险的URL，单击**查看详情**确认风险内容。
   - 消除风险后，单击**已处理**，完成处理。
   - 如果您对结果有异议，您可以单击**问题反馈**，将问题反馈给我们。在确认是问题后，我们将在算法层面进行优化改进。

5.RAM用户权限

5.1. 为RAM用户配置内容安全只读权限

内容安全支持通过RAM用户账号访问，但是RAM默认只提供内容安全的管理权限，您只能授权RAM用户完全管理内容安全。如果您希望仅授权RAM用户只读访问内容安全，不能进行写操作（例如OSS违规扫描配置、操作控制台扫描结果等），您可以使用自定义权限策略进行授权。

前提条件

已创建RAM用户。更多信息，请参见创建RAM用户。

操作步骤

1. 登录阿里云访问控制控制台。
2. 自定义内容安全的只读权限策略。
   i. 在左侧导航栏，选择**权限管理 > 权限策略管理**。
   ii. 单击**创建权限策略**。
   iii. 在**创建权限策略**页面的**脚本编辑**页签，并填入以下脚本内容。

```
{
    "Version": "1",
    "Statement": [{
        "Action": [
            "yundun-greenweb:List*",
            "yundun-greenweb:Get*",
            "yundun-greenweb:Describe*",
            "yundun-greenweb:Query*"
            ],
        "Resource": "*",
        "Effect": "Allow"
    }]
}
```

   iv. 单击**下一步**。
   v. 输入策略的**名称**，然后单击**确定**。

   自定义权限策略新建成功。
3. 为RAM用户授予内容安全的只读权限。

i. 在左侧导航栏，选择**身份管理 > 用户**。

ii. 在**用户**页面，定位到要操作的用户，单击其**操作**列下的**添加权限**。

iii. 在**添加权限**面板，在**自定义策略**页签下为用户选择步骤2中新建的内容安全只读权限，然后单击**确定**。

iv. 确认授权结果，并单击**完成**。

授权成功。

4. （可选）如果在获得内容安全只读权限前，RAM用户已拥有内容安全的管理权限（对应系统权限策略AliyunYundunGreenWebFullAccess），您可以为RAM用户移除内容安全的管理权限。

i. 在**用户**页面，单击目标**用户名称**。

ii. 在用户详情页面，单击**权限管理**。

iii. 在**个人权限**页签下，定位到**AliyunYundunGreenWebFullAccess**系统权限策略，单击其**操作**列下**移除权限**。然后单击**确定**。

> **说明** 如果RAM用户通过用户组继承了AliyunYundunGreenWebFullAccess权限（可以在**继承用户组的权限**页签下确认），则您必须通过用户组移除该权限或将当前RAM用户从用户组中移出。更多信息，请参见为用户组移除权限、移出用户组成员。

成功移除权限。

5.2. 使用RAM用户调用内容安全API

内容检测API支持您通过RAM用户（即阿里云子账号）的方式进行调用。要使用RAM用户调用内容检测API，您需要创建RAM用户并完成授权。

操作步骤

1. 登录RAM控制台，创建RAM用户， 并选择生成AccessKeyID和AccessKeySecret。创建后请妥善保存该AccessKeyID和AccessKeySecret，后续使用SDK时需要提供。

   更多关于创建RAM用户的操作，请参见创建RAM用户。

2. 完成RAM用户授权。只有完成RAM用户授权，您的RAM用户才能够调用相关API。您需要向RAM用户授权以下系统策略权限：`AliyunYundunGreenWebFullAccess`。

   更多关于RAM用户授权的操作，请参见为RAM用户授权。

3. 使用RAM用户调用内容检测API。

   更多内容，请参见API概览。