ALIBABA CLOUD

Alibaba Cloud

内容安全

User Guide

Document Version: 20200910

Alibaba Cloud

Legal disclaimer

Alibaba Cloud reminds you to carefully read and fully understand the terms and conditions of this legal disclaimer before you read or use this document. If you have read or used this document, it shall be deemed as your total acceptance of this legal disclaimer.

1. You shall download and obtain this document from the Alibaba Cloud website or other Alibaba Cloud-authorized channels, and use this document for your own legal business activities only. The content of this document is considered confidential information of Alibaba Cloud. You shall strictly abide by the confidentiality obligations. No part of this document shall be disclosed or provided to any third party for use without the prior written consent of Alibaba Cloud.

2. No part of this document shall be excerpted, translated, reproduced, transmitted, or disseminated by any organization, company or individual in any form or by any means without the prior written consent of Alibaba Cloud.

3. The content of this document may be changed because of product version upgrade, adjustment, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and an updated version of this document will be released through Alibaba Cloud-authorized channels from time to time. You should pay attention to the version changes of this document as they occur and download and obtain the most up-to-date version of this document from Alibaba Cloud-authorized channels.

4. This document serves only as a reference guide for your use of Alibaba Cloud products and services. Alibaba Cloud provides this document based on the "status quo", "being defective", and "existing functions" of its products and services. Alibaba Cloud makes every effort to provide relevant operational guidance based on existing technologies. However, Alibaba Cloud hereby makes a clear statement that it in no way guarantees the accuracy, integrity, applicability, and reliability of the content of this document, either explicitly or implicitly. Alibaba Cloud shall not take legal responsibility for any errors or lost profits incurred by any organization, company, or individual arising from download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, take responsibility for any indirect, consequential, punitive, contingent, special, or punitive damages, including lost profits arising from the use or trust in this document (even if Alibaba Cloud has been notified of the possibility of such a loss).

5. By law, all the contents in Alibaba Cloud documents, including but not limited to pictures, architecture design, page layout, and text description, are intellectual property of Alibaba Cloud and/or its affiliates. This intellectual property includes, but is not limited to, trademark rights, patent rights, copyrights, and trade secrets. No part of this document shall be used, modified, reproduced, publicly transmitted, changed, disseminated, distributed, or published without the prior written consent of Alibaba Cloud and/or its affiliates. The names owned by Alibaba Cloud shall not be used, published, or reproduced for marketing, advertising, promotion, or other purposes without the prior written consent of Alibaba Cloud. The names owned by Alibaba Cloud include, but are not limited to, "Alibaba Cloud", "Aliyun", "HiChina", and other brands of Alibaba Cloud and/or its affiliates, which appear separately or in combination, as well as the auxiliary signs and patterns of the preceding brands, or anything similar to the company names, trade names, trademarks, product or service names, domain names, patterns, logos, marks, signs, or special descriptions that third parties identify as Alibaba Cloud and/or its affiliates.

6. Please directly contact Alibaba Cloud for any errors of this document.

Document conventions

Style	Description	Example
Danger	A danger notice indicates a situation that will cause major system changes, faults, physical injuries, and other adverse results.	Danger: Resetting will result in the loss of user configuration data.
Warning	A warning notice indicates a situation that may cause major system changes, faults, physical injuries, and other adverse results.	Warning: Restarting will cause business interruption. About 10 minutes are required to restart an instance.
Notice	A caution notice indicates warning information, supplementary instructions, and other content that the user must understand.	Notice: If the weight is set to 0, the server no longer receives new requests.
Note	A note indicates supplemental instructions, best practices, tips, and other content.	Note: You can use Ctrl + A to select all files.
>	Closing angle brackets are used to indicate a multi-level menu cascade.	Click **Settings**> **Network**> **Set network type**.
Bold	Bold formatting is used for buttons , menus, page names, and other UI elements.	Click **OK**.
`Courier font`	Courier font is used for commands	Run the `cd /d C:/window` command to enter the Windows system folder.
Italic	Italic formatting is used for parameters and variables.	`bae log list --instanceid` *Instance_ID*
[] or [a|b]	This format is used for an optional value, where only one item can be selected.	`ipconfig [-all|-t]`
{} or {a|b}	This format is used for a required value, where only one item can be selected.	`switch {active|stand}`

Table of Contents

1.Content Moderation API

1.1. Customize policies for machine-assisted moderation

Content Moderation uses the default moderation policy of Alibaba Cloud for machine-assisted moderation. If you find that the default moderation policy cannot meet your business needs during testing, you can customize policies for machine-assisted moderation based on the templates provided by Content Moderation. This topic describes how to customize a policy for machine-assisted moderation.

Context

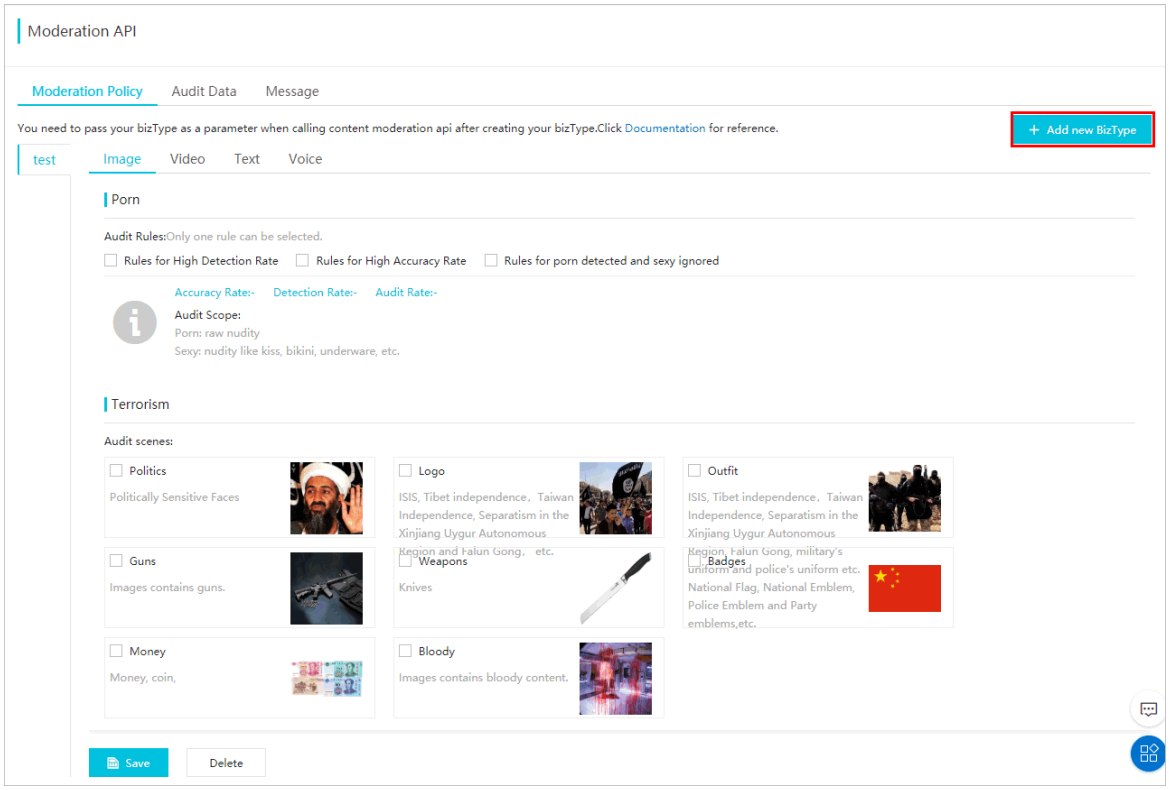
Currently, Content Moderation provides templates for you to customize policies for moderating images for pornographic content and terrorist content.

Before you customize a policy for machine-assisted moderation, get familiar with the following concepts:

- bizType: the business scenario. Moderation policies are customized based on the bizType setting. Each bizType setting corresponds to a moderation policy. If you do not customize moderation policies, the default bizType setting and corresponding default moderation policy are used. After you add a bizType setting, you can use the bizType setting to specify the bizType parameter in an API request for content moderation. In this case, the corresponding moderation policy takes effect.
- Accuracy rate: the ratio of the number of objects that are detected as violations by machine-assisted moderation and tagged as violations by human review to the number of objects that are detected as violations by machine-assisted moderation.
- Detection rate: the ratio of the number of objects that are detected as violations by machine-assisted moderation and tagged as violations by human review to the number of objects that are tagged as violations by human review.
- Review rate: the ratio of the number of objects that are detected as suspected violations by machine-assisted moderation to the total number of moderated objects.

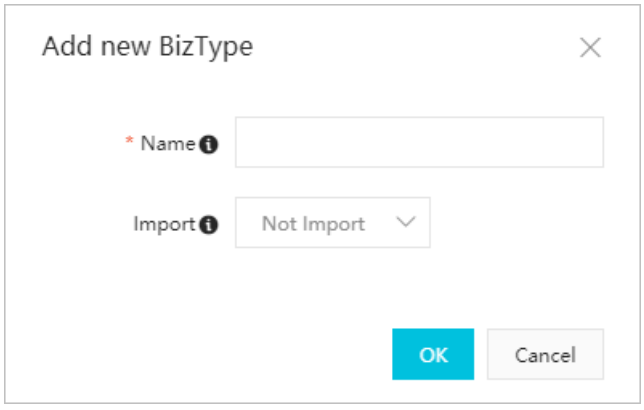
Procedure

1. Log on to the Alibaba Cloud Content Moderation console.
2. In the left-side navigation pane, choose **Settings** > **Moderation API**. The **Moderation Policy** tab of the Moderation API page appears.
3. On the Moderation Policy tab, click the Create bizType icon for the first time. Click **Add new BizType** if you have created a bizType setting.



4. In the **Add new BizType** dialog box that appears, set relevant parameters and click **OK**. The following table describes the parameters.

Parameter	Description
Name	The name of the bizType setting. The name can contain digits, letters, and underscores (_). It can be up to 32 characters in length.
Import	You can import an existing bizType setting.



The bizType setting is created. You can view the created bizType setting in the left-side list.

5. Click the tab of the created bizType setting and customize a moderation policy. The moderation policy template varies with the moderation scenario. For more information, see the templates in the console. Currently, you can only customize the moderation policy to moderate images for pornographic content and terrorist content. The following table describes the configuration of the moderation policy.

Moderation object	Moderation scenario	Moderation policy	Description
Image	Pornography detection	The moderation rule for pornography detection. You can select only one rule. Valid values: ○ Rules for High Detection Rate ○ Rules for High Accuracy Rate ○ Rules for porn detected and sexy ignored	When you call an API operation to moderate images for pornographic content, set the scenes parameter to *porn* in the API request. In this case, the moderation policy takes effect. For more information, see Moderate images synchronously.
Image	Terrorist content detection	The scenarios to be controlled for terrorist content detection. You can select one or more scenarios. Valid values: ○ Politics ○ Logo ○ Outfit ○ Guns ○ Weapons ○ Badges ○ Money ○ Bloody	When you call an API operation to moderate images for terrorist content, set the scenes parameter to *terrorism* in the API request. In this case, the moderation policy takes effect. For more information, see Moderate images synchronously.

6. After the configuration is completed, click **Save**.

Result

The moderation policy is customized.

What's next

Inform developers of the bizType setting of the moderation policy to be applied. Then, the developers can use the bizType setting to specify the bizType parameter when they follow the API reference to call the Content Moderation API. In this case, the custom moderation policy corresponding to the bizType setting is used during machine-assisted moderation.

For example, you want to call a synchronous operation to moderate images. If you set the bizType parameter in the API request to the name of the bizType setting you specified in the console, the corresponding moderation policy is applied to image moderation.

1.2. Manage sample libraries

1.2.1. Manage custom image libraries

Content Moderation bases its moderation service on the global image library of Alibaba Cloud by default, which can meet most of the moderation needs. To meet specific business needs, Content Moderation also supports custom image libraries. You can manage images to be blocked, passed, or reviewed separately in different custom image libraries to handle emergencies.

Context

Custom image libraries can be divided into blacklists, whitelists, and review lists based on the purposes. If custom image libraries are used for moderation, the images to be moderated that hit samples in custom image libraries are tagged with the corresponding category of moderation results.

- The moderation result for images that hit samples in blacklists is block.
- The moderation result for images that hit samples in whitelists is pass.
- The moderation result for images that hit samples in review lists is review.

Custom image libraries consist of feedback-based image libraries and user-created image libraries:

- Feedback-based image libraries are automatically created to accommodate the images that are reviewed. By default, you can use feedback-based image libraries to moderate images in all moderation scenarios of the same type. You can manage the images in feedback-based image libraries. However, you cannot perform operations on feedback-based image libraries, such as disabling or deleting a feedback-based image library.
- User-created image libraries are created by you to moderate images in a specific or a type of moderation scenario. You can manage the images in user-created image libraries and perform operations on user-created image libraries.

> **Note** You can create up to 10 user-created image libraries and add up to 10,000 images to each user-created image library.

When using the Content Moderation API, you can use custom image libraries to moderate images or videos for pornographic content, terrorist content, ad violations, and undesirable scenes.

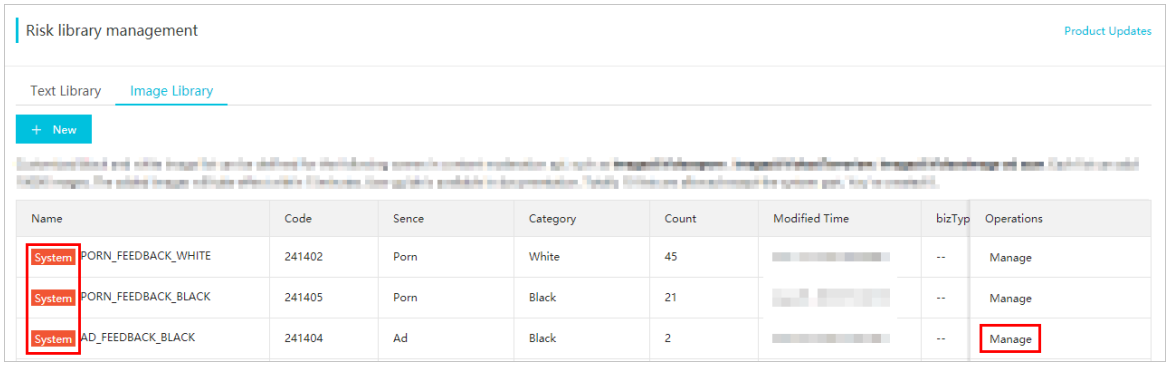
This topic describes how to manage custom image libraries for the Content Moderation API in the Alibaba Cloud Content Moderation console. In addition to operations in the console, you can also manage custom image libraries by using the API or SDK. For more information, see the following methods:

- Use the API to manage custom image libraries
- Use the Java SDK to manage custom image libraries

Manage feedback-based image libraries

1. Log on to the Alibaba Cloud Content Moderation console.
2. In the left-side navigation pane, choose **Moderation API > Risk library management**.
3. On the Risk library management page that appears, click the **Image Library** tab. On this tab, find the target feedback-based image library and click **Manage** in the Operations column. The **Image Library** tab lists all custom image libraries. The libraries marked with **System** and named in SCENARIO_FEEDBACK_WHITE or SCENARIO_FEEDBACK_BLACK format are feedback-based image libraries. For example, the **PORN_FEEDBACK_BLACK** library is a blacklist that consists of samples added by the system and is used to moderate images for pornographic

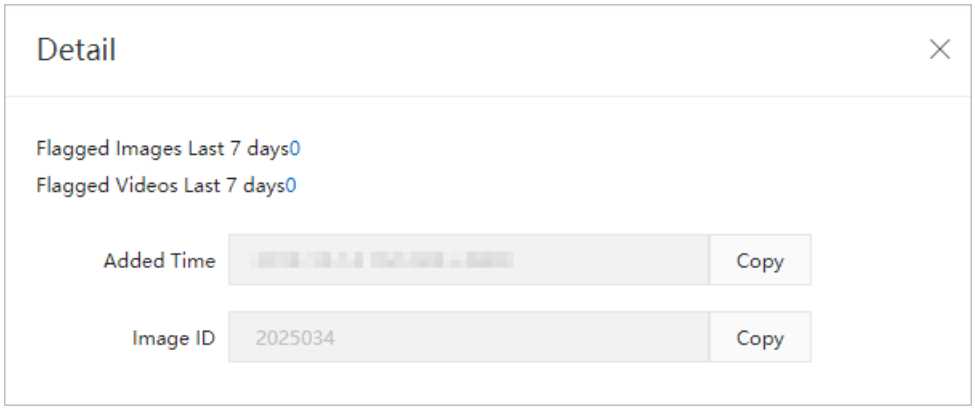
content.



4. On the **Manage Image Library** page that appears, perform the following operations as needed:

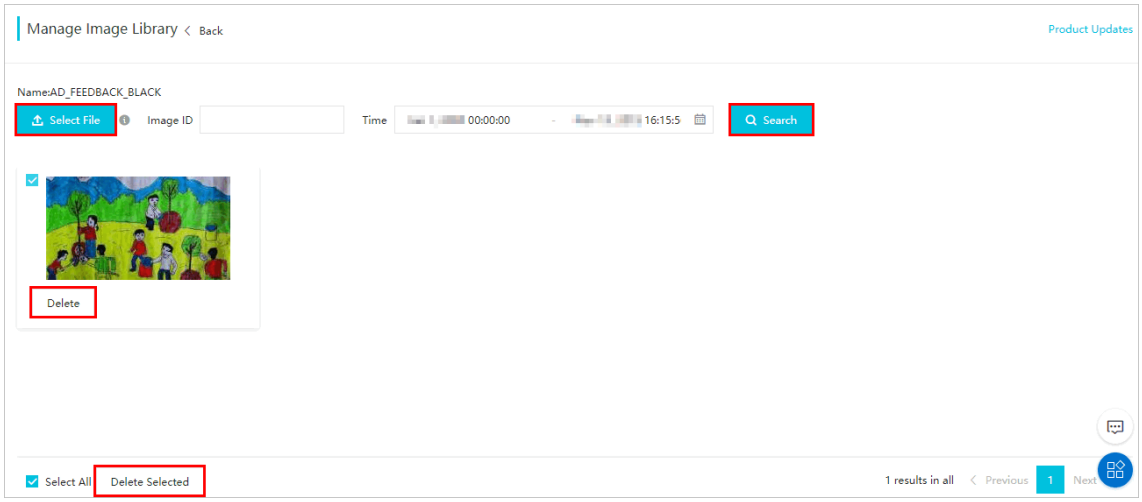
> **Note** You can add and delete images. The operations take effect in 15 minutes.

- Query images by setting the **Image ID** and **Time** parameters.
- Click an image. In the **Detail** dialog box that appears, view the respective numbers of images and videos that hit the image, the time when the image was added, and the ID of the image.



- Click **Select File** and upload images to the current image library.

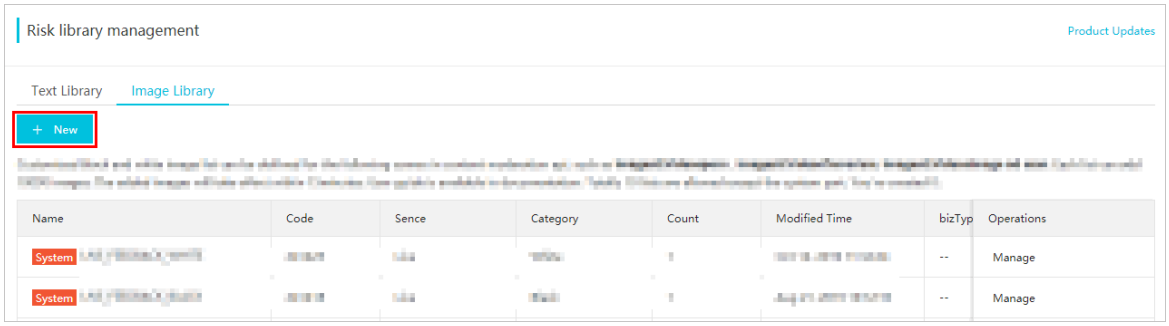
> **Note** You can upload images in PNG, JPG, JPEG, or BMP format. The size of an image cannot exceed 5 MB. You can upload up to 20 images at a time.



- Click **Delete** to delete an image. Alternatively, select multiple images and click **Delete Selected** at the bottom of the page.

Create and manage custom image libraries

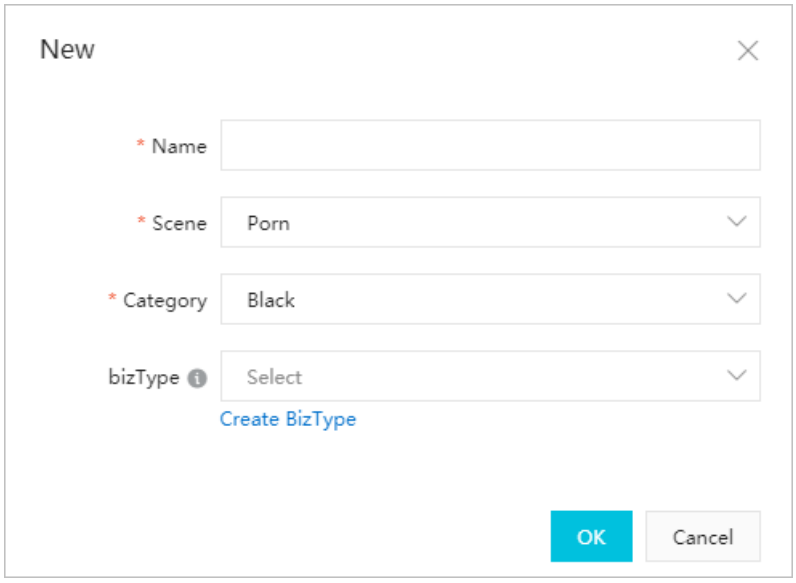
1. Log on to the Alibaba Cloud Content Moderation console.
2. In the left-side navigation pane, choose **Moderation API > Risk library management**. On the Risk library management page, click the **Image Library** tab.
3. On the Image Library tab, click **New**.



4. In the **New** dialog box that appears, set relevant parameters and click **OK**. The following table describes the parameters.

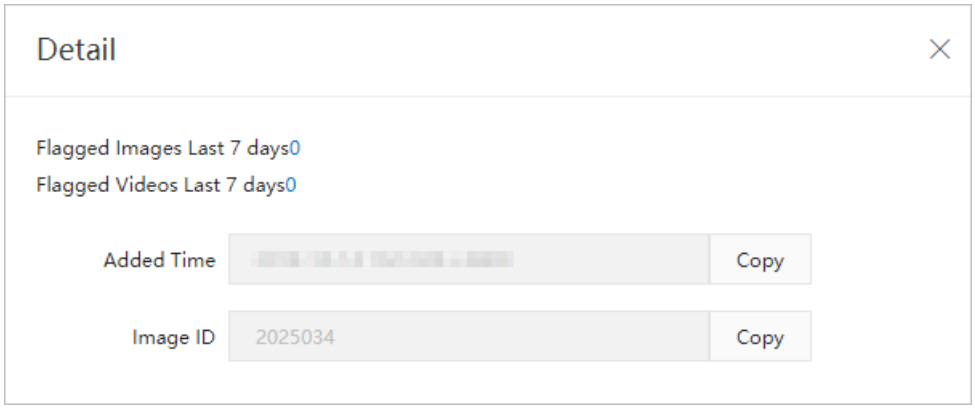
Parameter	Description
Name	The name of the custom image library. The name can be up to 64 characters in length. We recommend that you give the image library a readable name.

Parameter	Description
Scene	The moderation scenario to which the custom image library applies. Valid values: ○ **Porn**: moderates images or videos for pornographic content, where the value of the scenes parameter contains porn in API requests. ○ **Terrorism**: moderates images or videos for terrorist content, where the value of the scenes parameter contains terrorism in API requests. ○ **Ad**: moderates images or videos for ad violations, where the value of the scenes parameter contains ad in API requests. ○ **Live**: moderates images or videos for undesirable scenes, where the value of the scenes parameter contains live in API requests.
Category	The category of moderation results returned based on the custom image library. Valid values: ○ **Black**: If the images to be moderated hit samples in the image library, the machine-assisted moderation result returns the suggestion parameter with a value of block. ○ **Review**: If the images to be moderated hit samples in the image library, the machine-assisted moderation result returns the suggestion parameter with a value of review. ○ **White**: If the images to be moderated hit samples in the image library, the machine-assisted moderation result returns the suggestion parameter with a value of pass.
bizType	The business scenario to which the custom image library applies. You can specify different image libraries in API operations to meet business needs. For example, you can use the bizType parameter to specify the image library to be applied in a specific moderation scenario. The bizType parameter takes effect in the following ways: ○ If the bizType parameter in a moderation request is set to A, the image libraries of which the bizType parameter is set to A are used for moderation. These image libraries must be enabled. ○ In other cases, all enabled image libraries are used for moderation. Set this parameter as needed. We recommend that you submit a ticket to Alibaba Cloud to seek technical support.



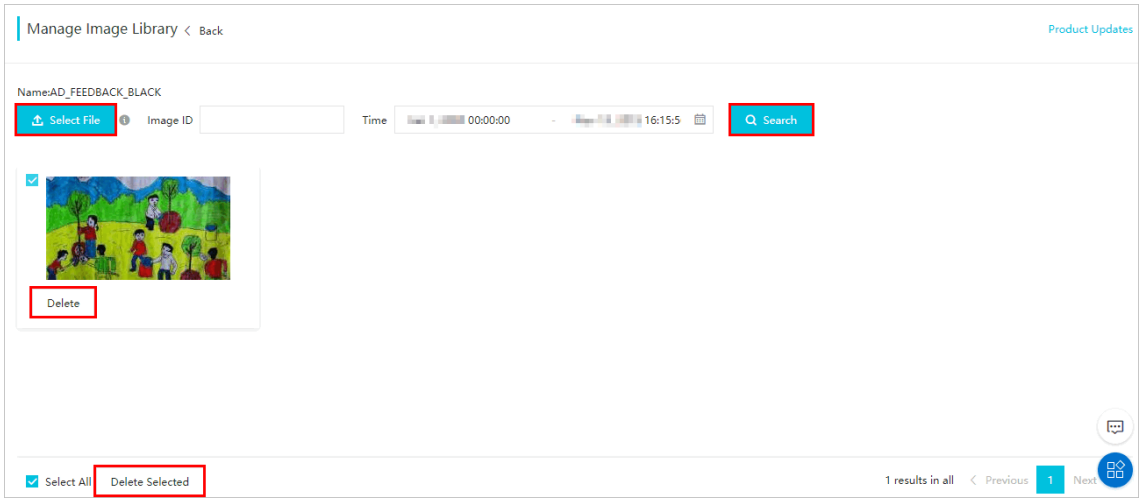
The custom image library is created. The status of the created image library is enabled by default.

5. Manage images in the created image library. Return to the Image Library tab. Find the target image library and click **Manage** in the **Operations** column. You can manage images on the **Manage Image Library** page that appears.
   - Query images by setting the **Image ID** and **Time** parameters.
   - Click an image. In the **Detail** dialog box that appears, view the respective numbers of images and videos that hit the image, the time when the image was added, and the ID of the image.



   - Click **Select File** and upload images to the current image library.

> **Note** You can upload images in PNG, JPG, JPEG, or BMP format. The size of an image cannot exceed 5 MB. You can upload up to 20 images at a time.



- Click **Delete** to delete an image. Alternatively, select multiple images and click **Delete Selected** at the bottom of the page.

6. Delete, modify, or disable an image library. Return to the Image Library tab. Select the target image library and click **Delete**, **Modify**, or **Disable** in the **Operations** column to perform the corresponding operation.

1.2.2. Manage custom text libraries

Content Moderation bases its moderation service on the global text library of Alibaba Cloud by default, which can meet most of the moderation needs. To meet specific business needs, Content Moderation also supports custom text libraries. You can manage text to be blocked, passed, or reviewed separately in different custom text libraries to handle emergencies.

Context

> **Notice** To seek technical support, you can submit a ticket to Alibaba Cloud. We recommend that you do not add terms unless necessary. Otherwise, the accuracy of moderation results cannot be guaranteed due to incorrect matches.

Custom text libraries consist of feedback-based text libraries and user-created text libraries:

- Feedback-based text libraries are automatically created to accommodate the text that is reviewed. For more information, see Review data. By default, you can use feedback-based text libraries to moderate text in all moderation scenarios of the same type. You can manage the text in feedback-based text libraries. However, you cannot perform operations on feedback-based text libraries, such as disabling or deleting a feedback-based text library.
- User-created text libraries are created by you to moderate text in a specific or a type of moderation scenario. You can manage the text in user-created text libraries and perform operations on user-created text libraries.

> **Note** You can create up to 10 user-created text libraries.

When using the Content Moderation API, you can apply custom text libraries to ad violation detection and text anti-spam.

This topic describes how to manage custom text libraries for the Content Moderation API in the Alibaba Cloud Content Moderation console. In addition to operations in the console, you can also manage custom text libraries by using the API or SDK. For more information, see the following methods:

- Use the API to manage custom text libraries
- Use the Java SDK to manage custom text libraries

Text types

The text in custom text libraries consists of terms and text patterns.

- Terms

  Terms are designed to moderate words in text. If a sentence or a piece of text contains a certain term, the term is hit. You can add different terms for different business scenarios.

  In Content Moderation, you can apply term-based moderation to ad violation detection and text anti-spam. For more information about relevant parameters, see the parameter description of moderation operations in different scenarios.

  You can add the AND (&) and NOT (~) logical operators in Chinese terms. For example:

  - The term "A&B" is added. If a piece of text contains both A and B, the term is hit.
  - The term "A~B" is added. If a piece of text contains A but does not contain B, the term is hit.

  > **Note** If you add both logical operators in a term, the AND (&) operator must be added before the NOT (~) operator. For example, you can add "A&B~C" as a term, but cannot add "A~C&B" as a term.

- Text patterns

  Text patterns are designed to compare the similarity between sentences or text. If two sentences or two pieces of text are partially different but express the same meaning, the two sentences or two pieces of text show a close similarity. Content Moderation can determine whether a piece of text has a close similarity to a text pattern in text pattern libraries. If the similarity reaches a certain degree, the text pattern is hit.

  You can apply text pattern libraries to text anti-spam. Content Moderation allows you to customize a blacklist, a whitelist, and a review list for text pattern libraries based on your business needs. The review list contains the text that needs human review. You can manage text patterns related to your business in text pattern libraries. In this case, the content that hits text patterns can be filtered out in text anti-spam.

Limits

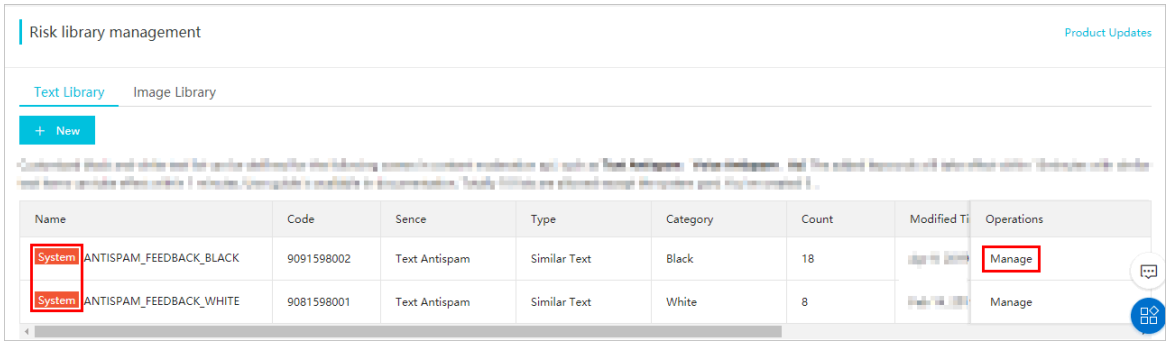
Type	Item	Limit
User-created text library	Quantity	Supports a maximum of 10 user-created text libraries.
User-created text library	Name length	Supports a maximum of 20 characters in length for each library name.

Type	Item	Limit
Term library	Term type	Supports Chinese characters and combinations of letters and digits. Currently, English words or phrases cannot be used as terms. Note Each combination of letters and digits is considered as a word during word-breaking.
Term library	Number of terms in a text library	Supports a maximum of 10,000 terms in a text library.
Term library	Term length	Supports a maximum of 50 characters in length for each term, including logical operators.
Term library	Encoding for Chinese terms	Supports UTF-8 encoding.
Term library	Term format	Excludes the following full-width and half-width special characters: at signs (@), number signs (#), dollar signs ($), percent signs (%), carets (^), asterisks (*), parentheses (), angle brackets (<>), forward slashes (/), question marks (?), commas (,), periods (.), semicolons (;), underscores (_), plus signs (+), hyphens (-), equal signs (=), single quotation marks ('), double quotation marks ("), spaces, and tabs.
Text pattern library	Text pattern library	Supports 10 to 4,000 characters in length for each text pattern. Note If the text added to the text library is too long, it may cause incorrect matches. We recommend that each text pattern be up to 200 characters in length. You can submit a ticket to seek technical support.
Text pattern library	Number of text patterns in a text library	Supports a maximum of 10,000 text patterns in a text library.
Text pattern library	Encoding	Supports UTF-8 encoding.
Text pattern library	Text content	Requires clear Chinese semantic features that can be extracted. If few semantic features can be identified from a text pattern, this text pattern is ignored. Note A text pattern that consists of meaningless letters, digits, or emoticons may be ignored.

Manage feedback-based text libraries

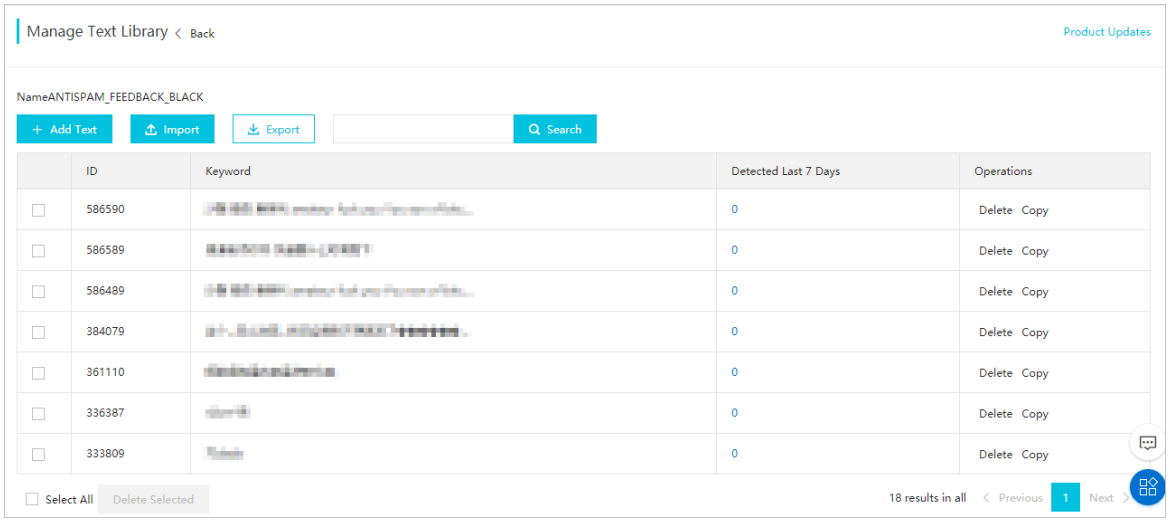
> **Note** Feedback-based text libraries contain text patterns rather than terms. You can manage custom text libraries whose text type is text pattern in the same way.

1. Log on to the Alibaba Cloud Content Moderation console.
2. In the left-side navigation pane, choose **Moderation API > Risk library management**. The **Text Library** tab of the Risk library management page appears.
   The **Text Library** tab lists all custom text libraries. The libraries marked with **System** and named in SCENARIO_FEEDBACK_WHITE or SCENARIO_FEEDBACK_BLACK format are feedback-based text libraries. For example, the **ANTISPAM_FEEDBACK_BLACK** library is a blacklist that consists of text patterns added by the system and is used for text anti-spam.



3. Find the target feedback-based text library and click **Manage** in the Operations column.

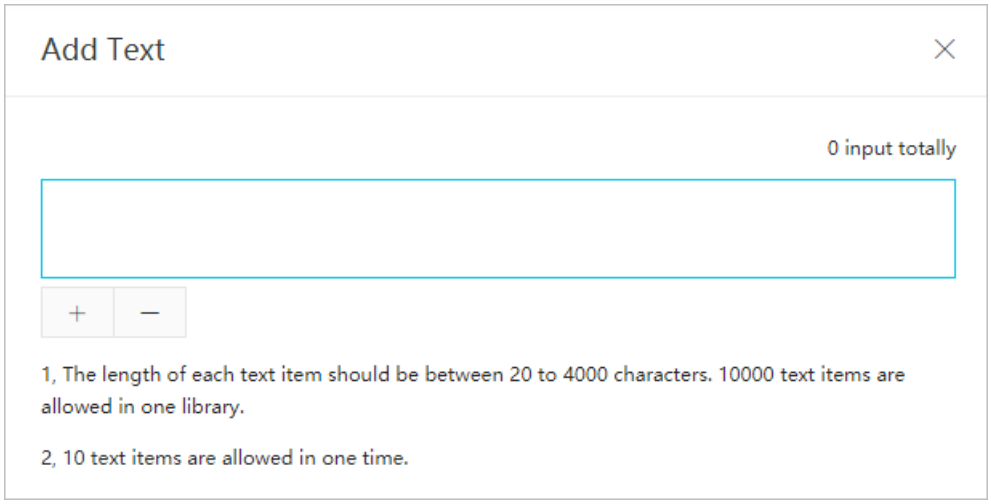
   The **Manage Text Library** page appears. This page lists all text patterns added to the library and displays the number of times that each text pattern is hit in the last seven days in the **Detected Last 7 Days** column, excluding the statistics on the current day.



4. On the **Manage Text Library** page, manage text patterns in the library.

   > **Note** You can add and delete text patterns. The operations take effect in 15 minutes.

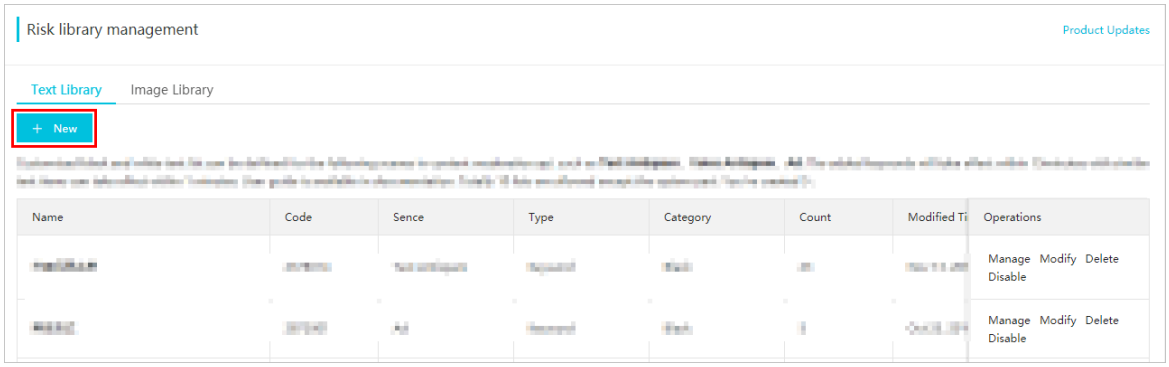
   - Click **Add Text** or **Import** and add text patterns as prompted.



- Select text patterns that you do not need and click **Delete Selected** at the bottom of the page. Alternatively, find a text pattern that you do not need and click **Delete** in the Operations column.

Create and manage custom text libraries

1. Log on to the Alibaba Cloud Content Moderation console.
2. In the left-side navigation pane, choose **Moderation API > Risk library management**. The **Text Library** tab of the Risk library management page appears.
3. On the Text Library tab, click **New**.

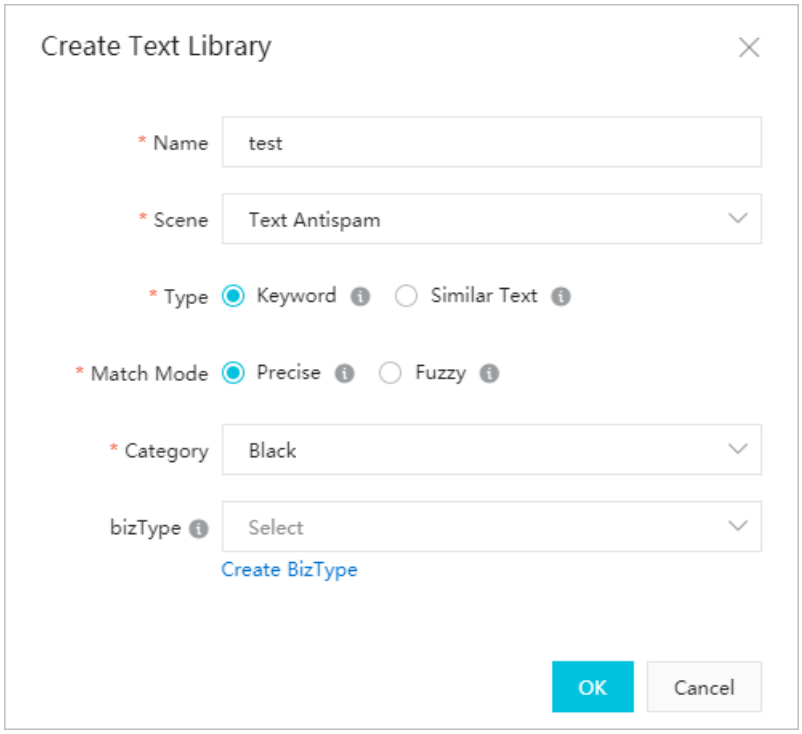


4. In the **Create Text Library** dialog box that appears, set relevant parameters and click **OK**. The following table describes the parameters.

Parameter	Description
Name	The name of the custom text library. You can set the same name for multiple text libraries. However, we recommend that you set a unique name for each text library.
Scene	The moderation scenario to which the custom text library applies. Valid values: ○ **Text Antispam**: text anti-spam where the value of the scenes parameter contains antispam in API requests ○ **Ad**: image moderation where the value of the scenes parameter contains ad in API requests

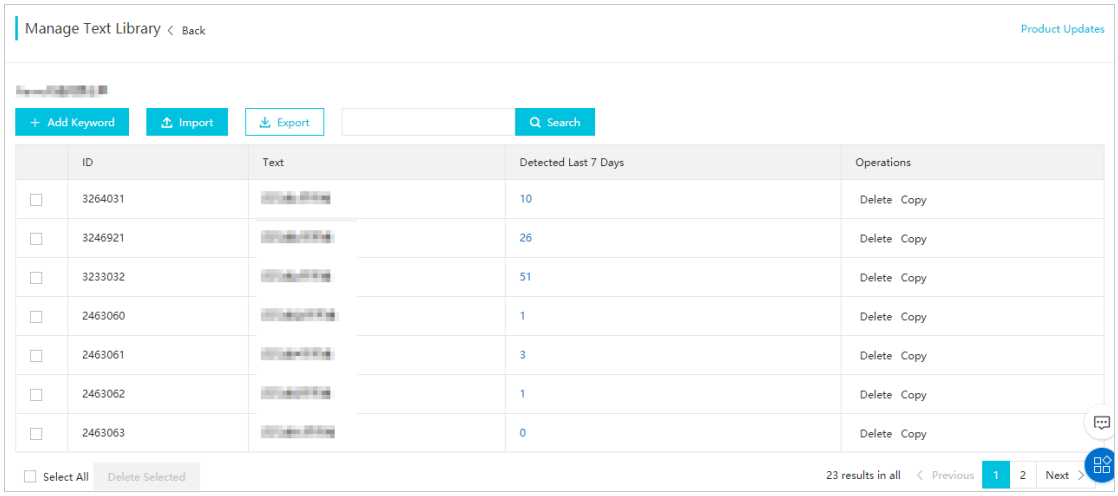
Parameter	Description
Type	The type of text in the custom text library. Valid values: ○ **Keyword**: matches the text to be moderated that contains terms. You can detect more risky text through terms. ○ **Similar Text**: matches the text to be moderated that is similar to text patterns at a certain probability. You can detect risky text more accurately through text patterns. **Note** You can set this parameter to Similar Text only when the **Scene** parameter is set to **Text Antispam**.
Match Mode	The match mode applied to the custom text library. This parameter is required when the **Type** parameter is set to **Keyword**. Valid values: ○ **Precise**: matches the text to be moderated that contains the same terms in the text library. ○ **Fuzzy**: preprocesses the text to be moderated and terms, and then matches the preprocessed text to be moderated that hits the preprocessed terms. The text to be moderated and terms are preprocessed in the following ways: ■ Convert uppercase letters to lowercase letters. For example, if the text to be moderated is "bitCoin", the term "bitcoin" is hit. ■ Convert traditional Chinese characters to simplified Chinese characters. ■ Convert similar words. For example, if the text to be moderated is "(2)", the term "2" is hit. **Note** The fuzzy mode is selected by default for text pattern libraries.

Parameter	Description
Category	The category of moderation results returned based on the custom text library. ○ Select **Keyword** for **Type**. Valid values: ■ **Black**: If the text to be moderated hits terms in the text library, the machine-assisted moderation result returns the suggestion parameter with a value of block. ■ **Review**: If the text to be moderated hits terms in the text library, the machine-assisted moderation result returns the suggestion parameter with a value of review. ■ **Ignore**: If the text to be moderated hits terms in the text library, the text is ignored and the machine-assisted moderation result returns the suggestion parameter with a value of pass. ○ Select **Similar Text** for **Type**. Valid values: ■ **Black**: If the text to be moderated hits text patterns in the text library, the machine-assisted moderation result returns the suggestion parameter with a value of block. ■ **Review**: If the text to be moderated hits text patterns in the text library, the machine-assisted moderation result returns the suggestion parameter with a value of review. ■ **White**: If the text to be moderated hits text patterns in the text library, the machine-assisted moderation result returns the suggestion parameter with a value of pass.
bizType	The business scenario to which the custom text library applies. You can specify different text libraries in API operations to meet business needs. For example, you can use the bizType parameter to specify the text library to be applied in a specific moderation scenario. The bizType parameter takes effect in the following ways: ○ If the bizType parameter in a moderation request is set to A, the text libraries of which the bizType parameter is set to A are used for moderation. These text libraries must be enabled. ○ In other cases, all enabled text libraries are used for moderation. Set this parameter as needed. We recommend that you submit a ticket to Alibaba Cloud to seek technical support.



The custom text library is created. You can view the created text library on the Text Library tab.

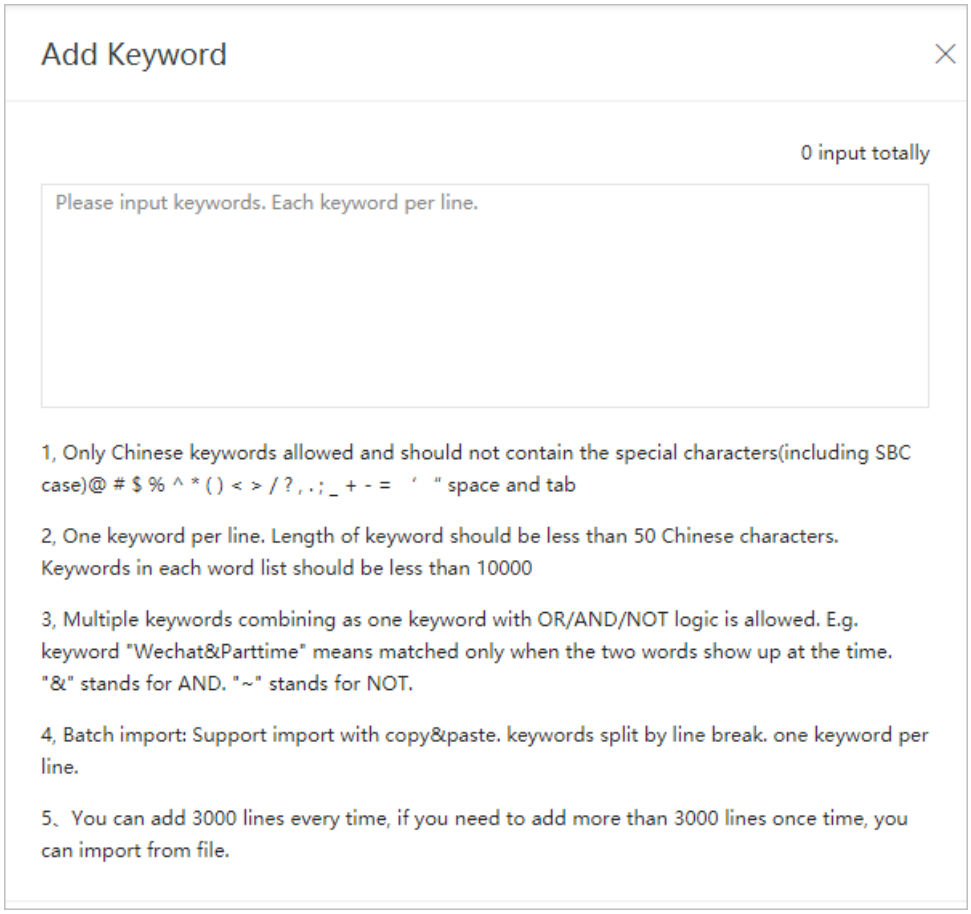
5. (Optional)If the text type of the created text library is term, follow these steps to manage terms. For more information about how to manage text patterns if the text type of the created text library is text pattern, see Manage feedback-based text libraries.
   i. Find the target text library whose text type is term and click **Manage** in the Operations column.
   The **Manage Text Library** page appears. This page lists all terms added to the library and displays the number of times that each term is hit in the last seven days in the **Detected Last 7 Days** column, excluding the statistics on the current day.



ii. On the **Manage Text Library** page, manage terms in the library.

> **Note** You can add and delete terms. The operations take effect in 15 minutes.

- Click **Add Keyword** or **Import** and add terms as prompted.



- Select terms that you do not need and click **Delete Selected** at the bottom of the page. Alternatively, find a term that you do not need and click **Delete** in the Operations column.

6. Delete, modify, or disable a custom text library. Return to the Text Library tab. Select the target text library and click **Delete**, **Modify**, or **Disable** in the Operations column to perform the corresponding operation.

1.3. Review data

The Alibaba Cloud Content Moderation console displays the result data detected by the Content Moderation API. You can review the machine-assisted moderation results based on your business scenarios. After human review, the Content Moderation API can moderate the same content based on your human review results. This topic describes how to review the machine-assisted moderation results.

Context

- By default, the Alibaba Cloud Content Moderation console displays only the result data whose machine-assisted moderation result is review or block on the Audit page for human

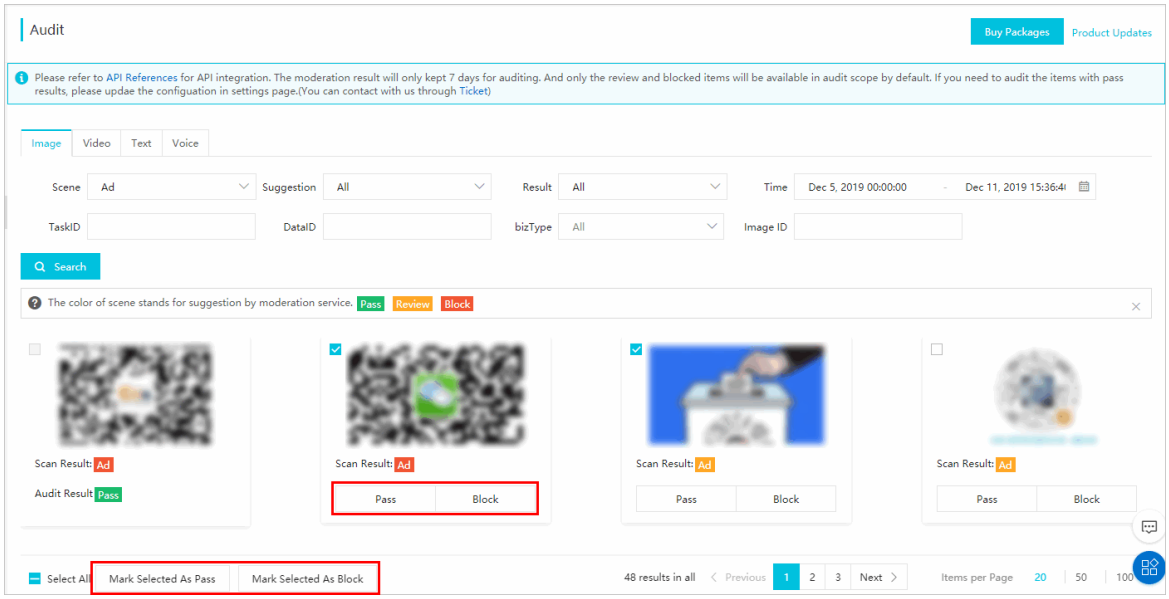
review. If you want to view the result data whose machine-assisted moderation result is pass on the Audit page, you can modify the settings in the console. For more information, see Select the data to be reviewed.

- During human review, you can review images, videos, and text. Only the images and text that you review can be automatically added to a sample library.
- The Alibaba Cloud Content Moderation console retains the machine-assisted moderation results for up to seven days. We recommend that you review the result data in a timely manner.

Procedure

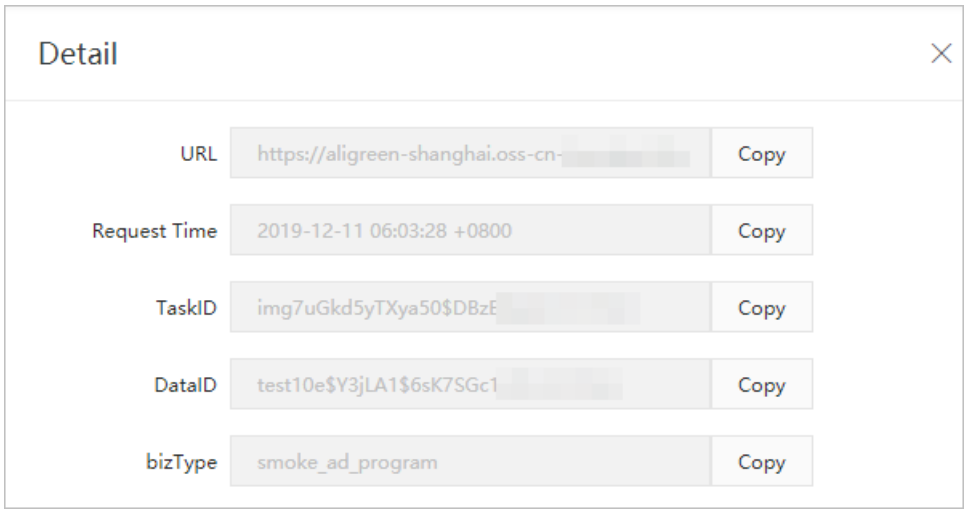
1. Log on to the Alibaba Cloud Content Moderation console.
2. In the left-side navigation pane, choose **Moderation API > Audit**.
3. On the Audit page that appears, click the **Image**, **Video**, or **Text** tab based on the type of data to be reviewed. The corresponding review tab appears.
4. Use the following methods to review the result data.

> **Note** This procedure shows how to review images. The methods for reviewing other types of data are similar.



- You can click **Pass** for the images that are detected as block or review but you consider as normal content.
- You can click **Block** for the images that are detected as review or pass but you consider as violations. Then, select **Porn**, **Terrorism**, **Ad**, or **Live** in the dialog box that appears.

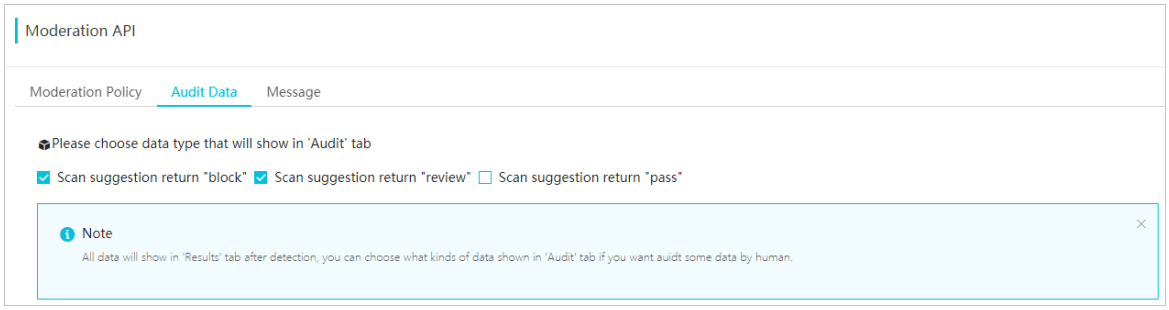
- You can select multiple images and review them at a time. Click **Mark Selected As Pass** or **Mark Selected As Block**.
- Click an image. You can view its details.



The moderation results of reviewed images and similar images are changed based on your human review results. The images with changed moderation results are added to corresponding libraries based on the human review results.

Select the data to be reviewed

1. Log on to the Alibaba Cloud Content Moderation console.
2. In the left-side navigation pane, choose **Settings** > **Moderation API**.
3. On the Moderation API page that appears, click the **Audit Data** tab and select the types of data to be reviewed. You can select the following data types:
   - **Scan suggestion return "block"**: This check box is selected by default.
   - **Scan suggestion return "review"**: This check box is selected by default.
   - **Scan suggestion return "pass"**



The modified settings immediately take effect.

1.4. Enable callback notification

Content Moderation can send asynchronous notifications to inform you of machine-assisted moderation results and your review results. If you want to use or integrate the results for your business purpose, you can enable callback notification. This topic describes how to enable callback notification for the Content Moderation API.

Context

Content Moderation supports callback notification for machine-assisted moderation results and human review results.

- Callback notification for machine-assisted moderation results: After a moderation request is processed, Content Moderation sends the machine-assisted moderation results to the specified HTTP callback URL by sending an HTTP POST request.
- Callback notification for human review results: After you review data or call feedback operations to rectify the machine-assisted moderation results, Content Moderation sends the human review results to the specified HTTP callback URL by sending an HTTP POST request. For more information, see Review data.

Concepts

The following table introduces the concepts related to callback notification.

Concept	Description
Callback URL	The public endpoint of your server that you specify in the Alibaba Cloud Content Moderation console. The callback URL must meet the following requirements: • Uses HTTP or HTTPS and is accessible from the Internet. • Supports the POST method. • Supports UTF-8-encoded data. • Supports receiving data in the `application` / `x-www-form-urlencoded` format. • Supports the checksum and content parameters in callback data.
Seed	The string that is used to verify whether the POST request is sent from Content Moderation to the specified HTTP callback URL.
Callback times	The number of times that Content Moderation pushes callback notifications to your server. If your server receives a callback notification, it sends HTTP status code 200 to Content Moderation. If your server fails to receive a callback notification, it sends other HTTP status codes to Content Moderation. After Content Moderation receives an HTTP status code other than 200, Content Moderation continues to push the callback notification until your server receives it. Content Moderation can push a callback notification repeatedly for up to 16 times.
Callback data	The content of the callback notification that Content Moderation sends to the specified callback URL. For more information about the parameters in the callback data, see Parameters in the callback data.

Parameters in the callback data

Parameter	Type	Description

Parameter	Type	Description
checksum	String	The string in the <UID> + <Seed> + <Content> format that is generated by the Secure Hash Algorithm 256 (SHA-256) algorithm. UID indicates the ID of your Alibaba Cloud account. You can query the ID in the Alibaba Cloud console. Note To prevent data tampering, you can use the SHA-256 algorithm to generate a string when your server receives a callback notification and verify the string against the received checksum parameter.
content	String	The JSON-formatted string to be parsed to the callback data in the JSON format. For more information about the callback data that is parsed from the content parameter, see the Content parameter description section of this topic.

Callback notification for machine-assisted moderation results

All asynchronous moderation operations of the Content Moderation API support callback notification, including asynchronous image moderation and asynchronous video moderation. For more information, see Moderate images asynchronously and Moderate videos asynchronously. If you call an asynchronous operation and need Content Moderation to return the moderation results, specify the callback and seed parameters in your moderation request. The callback parameter specifies the callback URL and the seed parameter specifies a string that is used to verify the callback notification request.

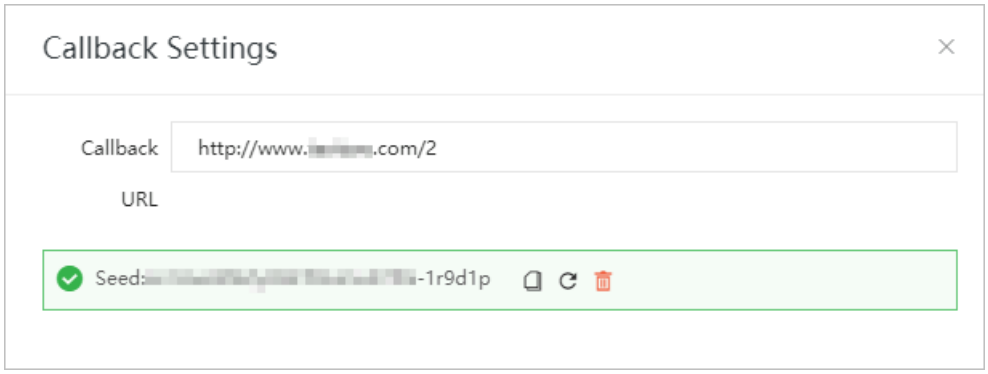
Note If you do not specify the preceding parameters when you call an asynchronous operation, you can only poll the asynchronous moderation results at regular intervals.

Procedure

1. Prepare the HTTP callback URL and verification string for receiving moderation results.
2. When you call an asynchronous operation for content moderation, specify the callback and seed parameters in your moderation request. For more information, see the parameter description of the relevant API operation.

Callback notification for human review results

1. Log on to the Alibaba Cloud Content Moderation console.
2. In the left-side navigation pane, choose **Settings** > **Moderation API**.
3. On the **Moderation API** page, click the **Message** tab.
4. On the **Message** tab, click **Settings** or **View**. If you have not specified a callback URL, the Settings button appears. If you have specified a callback URL, the View button appears.
5. In the **API Callback Notification Settings** pane, set **Callback URL** and click **OK**.
   After the callback URL is specified, the system automatically generates a value for the seed parameter. You can use the value of the seed parameter to verify whether a callback notification request is sent from Alibaba Cloud to the callback URL. Save and use the value of the seed parameter based on your business needs.



Content parameter description

After callback notification is enabled, Content Moderation sends the moderation results that are generated by the Content Moderation API in a callback notification to the specified callback URL. The callback notification contains the content parameter. The following tables describe the structure of the content parameter.

Structure of the content parameter

Parameter	Type	Required	Description
scanResult	JSON object	No	The machine-assisted moderation result. The structure of this parameter varies depending on the moderation object, such as images and videos. • For images, the structure is the same as that of the results parameter that is returned in response to synchronous operations for image moderation. For more information, see Moderate images synchronously. • For videos, the structure is the same as that of the results parameter that is returned in response to asynchronous operations for video moderation. For more information, see Moderate videos asynchronously.
auditResult	JSON object	No	The human review result that is generated by you. This parameter is returned only when human review is performed. For more information, see auditResult. Note This parameter is not returned if Content Moderation sends only machine-assisted moderation results.
humanAuditResult	JSON object	No	The human review result that is generated by the human review service of Alibaba Cloud. If you purchased the human review service of Alibaba Cloud, this parameter is returned to send the human review result. For more information, see humanAuditResult.

auditResult

Parameter	Type	Required	Description
suggestion	String	Yes	The suggestion that you provided during human review. Valid values: • *block* • *pass*
labels	JSON array	Yes	The tag that you set during human review. The returned value can be one or more tags. Valid values: • *porn* • *terrorism* • *ad* • *live*

humanAuditResult

Parameter	Type	Required	Description
suggestion	String	Yes	The suggestion that is provided by the human review service of Alibaba Cloud. Valid values: • *block* • *pass*
taskId	String	Yes	The ID of the moderation task. You can associate the human review result of the content with the corresponding machine-assisted moderation result based on the task ID.
dataId	String	Yes	The ID of the moderated content.
labels	String array	No	The tag that the human review service sets. The returned value can contain multiple tags. Note By default, this parameter is not returned. If this parameter is required, contact your business manager.

Example of the content parameter

```
{
  "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": [
       "Pornographic content",
       "Vulgar content"
    ],
     "taskId": "xxxxxx"
  }
}
```


1.5. Give feedback on samples

If the machine-assisted moderation results returned by Content Moderation do not meet your expectations, you can call feedback operations to give feedback on samples.

If you use the Content Moderation API on your own platform, you can also call feedback operations to give feedback on samples whose machine-assisted moderation results do not meet your expectations. Alibaba Cloud can add the feedback data to model training and correct the moderation results in later versions. The trained model can be more adaptive to your business scenarios.

Currently, Content Moderation allows you to give feedback on image samples, video samples, and text samples. For more information, see the API reference.

Automatically add samples to a custom image library

Model training requires sufficient samples and takes some time. You can enable Content Moderation to automatically add samples to a custom image library and correct the moderation results in real time.

For the feedback on image samples, you can manage feedback-based image libraries in the Alibaba Cloud Content Moderation console. The operations to manage feedback-based image libraries are the same as those to manage user-created image libraries, except that you cannot create or delete a feedback-based image library. For more information, see Manage custom image libraries.

To automatically add samples to a custom image library, you can set the label parameter when you call feedback operations:

- For normal images, set this parameter to *normal* to add these images to the whitelist.
- For risky images, set this parameter to other values to add these images to the blacklist. We recommend that you use values that indicate risks such as *porn*, *ad*, and *terrorism*.

1.6. View statistics

You can view the statistics about the Content Moderation API in the Alibaba Cloud Content Moderation console.

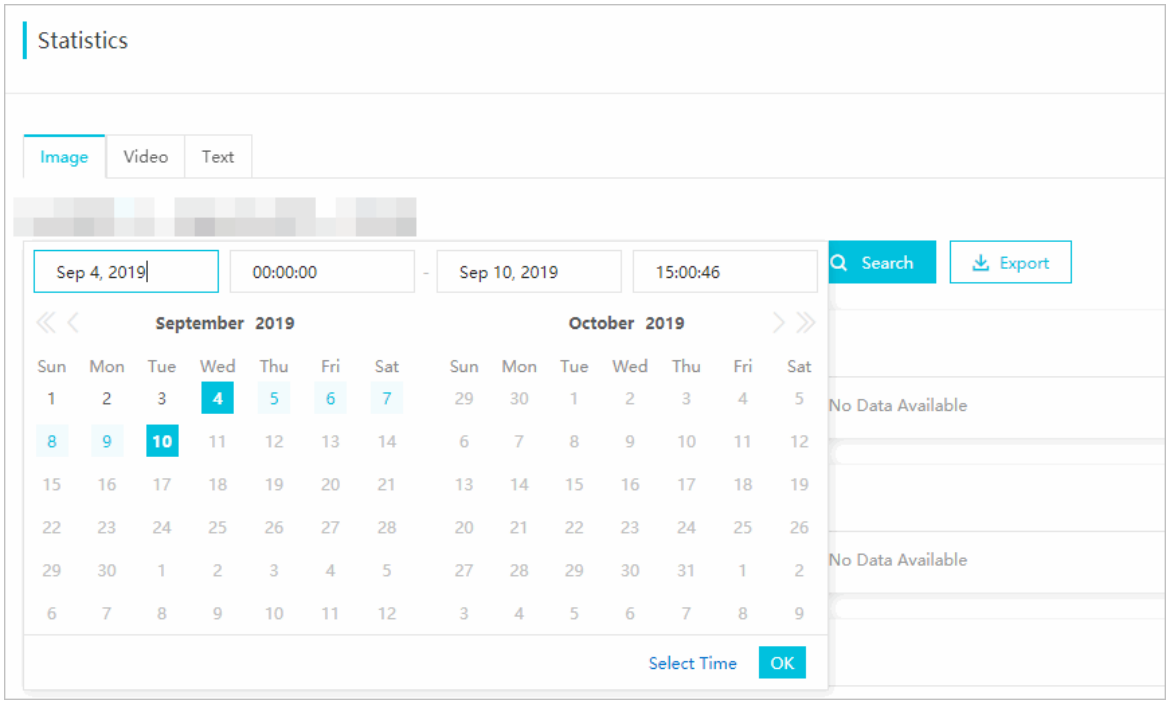
Context

The Alibaba Cloud Content Moderation console collects statistics on the Content Moderation API. You can query the number of times that the API is called to moderate images, videos, and text in the last year. You can also query the respective numbers of violations, suspected violations, and normal results in different moderation scenarios.

Procedure

1. Log on to the Alibaba Cloud Content Moderation console.
2. In the left-side navigation pane, choose **Moderation API > Statistics**.
3. On the Statistics page that appears, click the **Image**, **Video**, or **Text** tab based on the type of the data to be queried.
4. On the **Statistics** page, specify the time range that you want to query and click **Search**.

   You can query the data in the last year. You can query data for a maximum duration of a month.



5. View the statistics in different moderation scenarios and export the reports as needed.
   - Reports in the console
   - Exported reports

Reports in the console

For example, the **Porn** report shows the daily number of times that the Content Moderation API is called to moderate images for pornographic content, that is, the value of the scenes parameter contains porn in API requests. The report also shows the respective numbers of violations, suspected violations, and normal results. The legends are described as follows:

- **Block Count**: the number of moderation requests for which the value of the suggestion parameter is block in the returned moderation results.
- **Review Count**: the number of moderation requests for which the value of the suggestion parameter is review in the returned moderation results.
- **Pass Count**: the number of moderation requests for which the value of the suggestion parameter is pass in the returned moderation results.

The unit of the call volume varies with the moderation object:

- Image moderation: the number of images.
- Video moderation: the number of captured frames for a video or the duration of a video in units of minutes.
- Text moderation: the number of text entries.

Exported reports

You can export the reports in Excel format. The time range of each exported report is consistent with the time range that you set to query data. Each exported report in Excel format contains the statistics about the Content Moderation API that is called in a specific moderation scenario based on the value of the scenes parameter in API requests. The statistics are collected on a daily basis.

The following table describers the headers in the exported reports.

Header	Description	Unit
day	The date on which the API is called.	N/A
totalImageCount	The total number of images that are moderated.	Piece
blockImageCount	The number of images that contain violations.	Piece
reviewImageCount	The number of images that contain suspected violations.	Piece
passImageCount	The number of images that are normal.	Piece
totalVideoCount	The total number of videos that are moderated.	Piece
blockVideoCount	The number of videos that contain violations.	Piece
reviewVideoCount	The number of videos that contain suspected violations.	Piece
passVideoCount	The number of videos that are normal.	Piece
innerFrameCount	The number of video frames captured by Content Moderation.	Piece
outerFrameCount	The number of video frames submitted by users.	Piece
totalTextCount	The total number of text entries that are moderated.	Piece

Header	Description	Unit
blockTextCount	The number of text entries that contain violations.	Piece
reviewTextCount	The number of text entries that contain suspected violations.	Piece
passTextCount	The number of text entries that are normal.	Piece

1.7. Authorize a role to access ApsaraVideo for Media Processing

Content Moderation calls ApsaraVideo for Media Processing to capture video frames. In this way, data is accessed over the internal network to minimize the costs of Internet traffic. You must authorize a role to access ApsaraVideo for Media Processing. Then, Content Moderation can use the role to submit a video frame capture task. This topic describes how to create and authorize a role in the Resource Access Management (RAM) console.

Context

If you submit an asynchronous video moderation task and specify the video URL in the oss://xxxx format, Content Moderation can automatically capture frames for the video. To ensure automatic frame capture, you must authorize a role to access ApsaraVideo for Media Processing.

After you perform the operations described in this topic, you can achieve the following expected results:

- A RAM role is created for your Alibaba Cloud account and used by Content Moderation to call ApsaraVideo for Media Processing.

  For more information about RAM roles, see RAM role overview.
- The RAM role is granted the permission to access your Object Storage Service (OSS) buckets in the read-only mode.
- Video URLs to be submitted are generated in the oss://xxxx format.

> **Note** The method of submitting a video URL to specify the video to be moderated is available only for asynchronous video moderation tasks. For more information, see Moderate videos asynchronously.

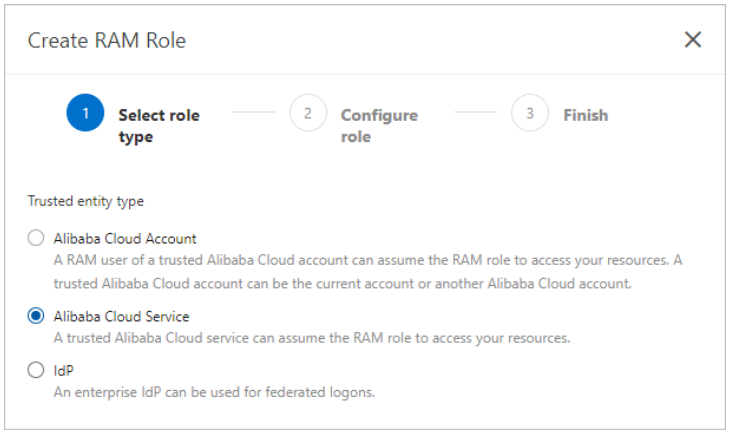
After a role is authorized, Content Moderation under your Alibaba Cloud account can use the role to call ApsaraVideo for Media Processing, access your OSS buckets, and then capture frames for videos in your OSS buckets.

Procedure

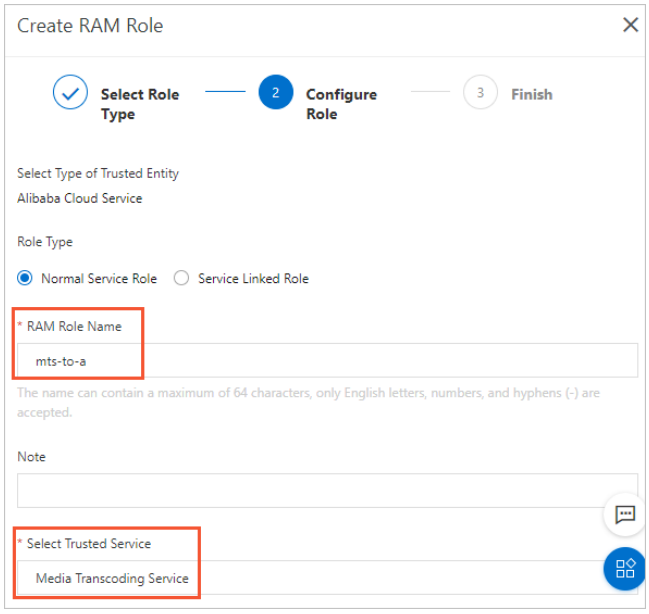
1. Create a RAM role.
   i. Log on to the RAM console.

ii. In the left-side navigation pane, click **RAM Roles**. On the RAM Roles page, click **Create RAM Role**.

iii. In the Create RAM Role right-side pane that appears, select **Alibaba Cloud Service** and click **Next**.

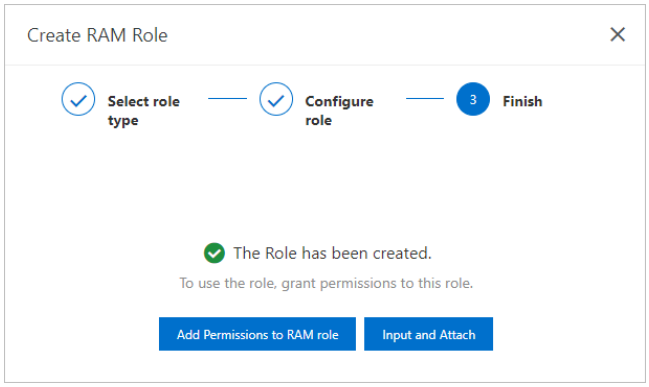


iv. Set the **RAM Role Name** parameter and select **Media Transcoding Service** from the **Select Trusted Service** drop-down list.



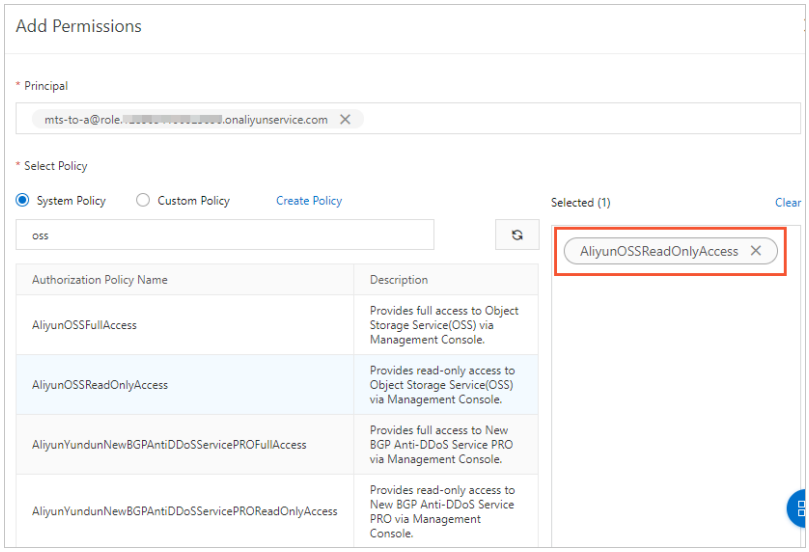
v. Click **OK**.

> **Note** This operation may require mobile phone-based authentication.



vi. After the role is created, click **Add Permissions to RAM Role** in the Create RAM Role right-side pane.

vii. In the **Add Permissions** right-side pane that appears, select the **AliyunOSSReadOnlyAccess** policy and click **OK**.

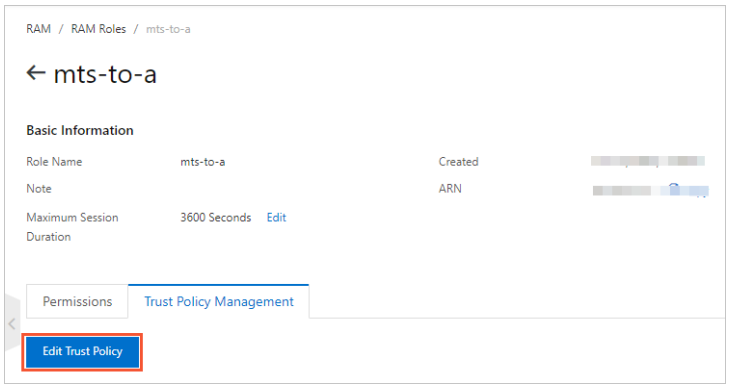


This policy grants the role the permission to access OSS buckets under your Alibaba Cloud account in the read-only mode.

2. Modify the trust policy.

i. On the **RAM Roles** page, find the created role and click the role name to view details.

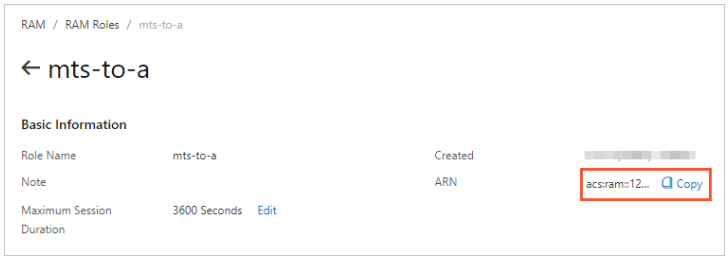
ii. Click the **Trust Policy Management** tab and then click **Edit Trust Policy**.



iii. In the **Edit Trust Policy** right-side pane that appears, replace the content of `"Service"` with `"1184847062244573@mts.aliyuncs.com"`. Click **OK**.

This operation specifies that the Alibaba Cloud account whose UID is 1184847062244573 assumes the created role to call ApsaraVideo for Media Processing.

3. On the role details page, view and copy the Alibaba Cloud resource name (ARN) in the **Basic Information** section.



4. Generate the video URL for a video to be moderated in an OSS bucket in the following format: `oss://arn@bucket.region/object`. Assume that you want to moderate the *video/bar.mp4* object in the *foo* OSS bucket that resides in the China (Shenzhen) region. The generated OSS URL of the video is `oss://acs:ram::xxxxxxxxxxxxxxxx:role/mts-to-a@foo.cn-shenzhen/video/bar.mp4`. `xxxxxxxxxxxxxxxx` is the 16-digit ID of your Alibaba Cloud account.

> **Note** The following regions support the preceding operation: China (Hangzhou), China (Shanghai), China (Beijing), and China (Shenzhen).

5. When you submit an asynchronous video moderation task, use the generated video URL.