Content Moderation

Product Introduction

Issue	01
Date	2025-07-03

HUAWEI TECHNOLOGIES CO., LTD.

Trademarks and Permissions

Notice

The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Security Declaration

Vulnerability

Huawei's regulations on product vulnerability management are subject to the *Vul. Response Process*. For details about this process, visit the following web page:
https://www.huawei.com/en/psirt/vul-response-process
For vulnerability information, enterprise customers can visit the following web page:
https://securitybulletin.huawei.com/enterprise/en/security-advisory

Contents

1 What Is Content Moderation?

Content Moderation adopts image, text, audio, audio stream, and video detection technologies that detect pornography and images and text violating related laws or regulations. This reduces potential business risks.

Malicious information, such as pornographic information bursts with the rapid development and information explosion of the Internet. Products with such information may annoy users and even lose user confidence.

Content Moderation provides services through open application programming interfaces (APIs). You can obtain the inference result by calling APIs. It helps you build an intelligent service system and improves service efficiency.

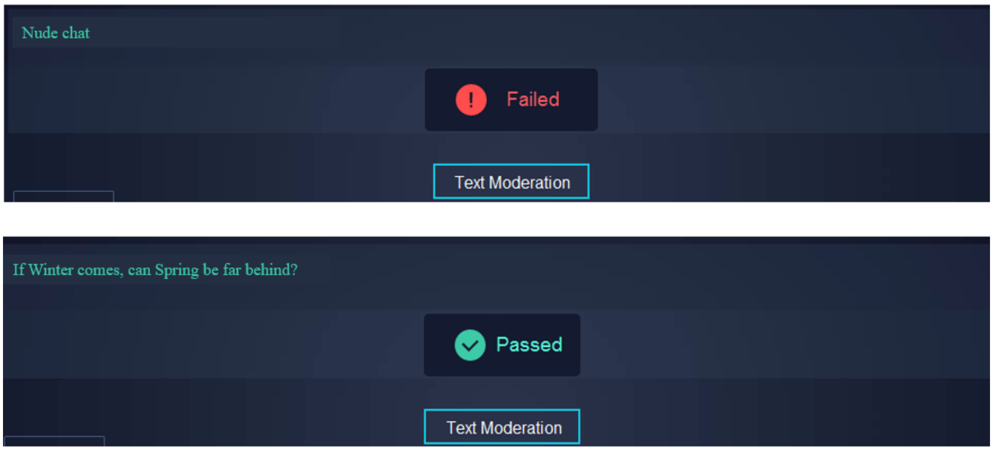
Image Moderation

Image Moderation uses the deep neural network (DNN) models to accurately identify pornography in images, protecting you from non-compliance risks.

Text Moderation

Text Moderation uses the AI-based text detection technology to detect non-compliant content, such as pornographic content, advertisements, and offensive content, and provide custom text moderation solutions.

Figure 1-1 Example of Text Moderation



Nude chat

If Winter comes, can Spring be far behind?

Audio Moderation

Audio Moderation adopts a leading speech recognition engine and an intelligent text detection model to accurately identify pornography and abuse in audio, greatly improving user experience.

Video Moderation

Video Moderation adopts advanced AI technologies to detect video images, sounds, and subtitles and accurately and efficiently identify pornography, violence, and advertisements, improving the content governance quality and efficiency.

Audio Stream Moderation

Audio Stream Moderation accurately identifies pornographic content, abuse, and advertisements in various scenarios to defend against content risks, improve audio stream review efficiency, and deliver better experience.

2 Advantages

Accurate Detection

The service uses deep learning technologies and a large number of sample libraries to help customers quickly and accurately detect non-compliant content and guarantee content security.

Versatile Functions

It can detect pornography, advertisements, and violent and terrorism-related content in text, images, audio, and videos.

Stability and Reliability

The service has been used by enterprises such as Huawei over the years and proves stable and reliable in complex scenarios.

High Efficiency

The service provides standard RESTful APIs and SDKs to facilitate use and integration while reducing labor and business costs.

3 Application Scenarios

Image Moderation

Application scenarios are as follows:

- Live video

  In the webcast scenario, thousands of channels are broadcasting concurrently, making manual review of broadcasting contents impossible. Moderation (Image) monitors, identifies, and flags live video channels with inappropriate, unwanted, or offensive content in real time.

  The advantages are as follows:

  - High accuracy: Moderation (Image) yields incredibly high levels of accuracy in detection with optimized deep learning algorithms.
  - High speed: Moderation (Image) responds to live video requests within 0.1 seconds.

- Online shopping

  Moderation (Image) identifies and warns of non-compliant images uploaded by sellers and users to prevent such images from being released, reducing manual reviews and non-compliance risks.

  The advantages are as follows:

  - High accuracy: Moderation (Image) yields high levels of accuracy in detection with optimized deep learning algorithms.
  - Rapid response: Moderation (Image) recognizes a single image within 0.1 seconds.

- Internet forum

  Moderation (Image) detects and flags non-compliant content to help you reduce your legal exposure.

  The advantages are as follows:

  - High accuracy: Moderation (Image) yields incredibly high levels of accuracy in detection with optimized deep learning algorithms.
  - Rapid response: Moderation (Image) recognizes a single image within 0.1 seconds.

Text Moderation

Application scenarios are as follows:

- E-commerce comment screening

  Moderation (Text) checks product comments on e-commerce websites and identifies non-compliant comments with pornographic elements and other types of events to ensure optimal user experience.

  The advantages are as follows:

  - High accuracy: Moderation (Text) yields incredibly high levels of accuracy in detection with optimized deep learning algorithms.
  - Rapid response: Moderation (Text) can respond within 0.1 seconds.

- User nickname review

  Intelligently reviews user registration information on websites and blocks nicknames that contain advertisements and pornographic elements.

  The advantages are as follows:

  - High accuracy: Moderation (Text) yields incredibly high levels of accuracy in detection with optimized deep learning algorithms.
  - Rapid response: Moderation (Text) can respond within 0.1 seconds.

- Media content review

  Automatically identifies contraband information in media content to avoid non-compliance risks in released articles.

  The advantages are as follows:

  - Fast update: Moderation (Text) continuously and quickly updates the dictionary to identify new non-compliant content in a timely manner.
  - High speed: Moderation (Text) can recognize non-compliant content within 0.1 seconds.

- Bullet Comment Review

  Moderation (Text) instantly reviews bullet comments that scroll across a screen to ensure webcast quality and reduce non-compliance risks.

  The advantages are as follows:

  - Large-scale dictionary: Moderation (Text) has a large-scale built-in dictionary with support for various matching rules.
  - Fast update: Moderation (Text) continuously and quickly updates the dictionary to identify new non-compliant content in a timely manner.

- Chat content review

  Moderation (Text) detects potential non-compliant information (such as offensive content, pornographic elements, and reactionary tendencies) in game chats in real time to purify the network environment.

  The advantages are as follows:

  - Large-scale dictionary: Moderation (Text) has a large-scale built-in dictionary with support for various matching rules.
  - Rapid response: Moderation (Text) can respond within 0.1 seconds.

Audio Moderation

Application scenarios are as follows:

- Online education

  Moderation (Audio) monitors audio teaching content and intelligently reviews violation scenarios such as pornography, violence, abuse, and advertisements.

  The advantages are as follows:

  - High accuracy: Moderation (Audio) uses optimized deep learning algorithms to improve the voice moderation accuracy in complex environments.
  - Recognition of special sounds: Moderation (Audio) provides models for recognizing special sounds, such as copulation calls, moaning, and sensitive voiceprints.

- Games/Social media apps

  Moderation (Audio) monitors the chat content and voice feeds in games and social media apps to reduce non-compliance risks.

  The advantages are as follows:

  - High accuracy: Moderation (Audio) uses optimized deep learning algorithms to improve the voice moderation accuracy in complex environments.
  - Recognition of special sounds: Moderation (Audio) provides models for recognizing special sounds, such as copulation calls, moaning, and sensitive voiceprints.

- Recording/Radio

  Audio Moderation monitors the audio data of content transmission and FM radio to reduce non-compliance risks.

  The advantages are as follows:

  - High accuracy: Moderation (Audio) uses optimized deep learning algorithms to improve the voice moderation accuracy in complex environments.
  - Recognition of special sounds: Moderation (Audio) provides models for recognizing special sounds, such as copulation calls, moaning, and sensitive voiceprints.

Video Moderation

Application scenarios are as follows:

- Video platforms/communities: Moderation (Video) accurately identifies non-compliant video content to help platforms and communities avoid risks.
  - Comprehensive detection: Moderation (Video) provides a multi-modal comprehensive review solution to parse the images, sounds, and text in videos.
  - Various types of video formats: Moderation (Video) supports video formats such as AVI, FLV, MP4, MPG, WMV, MOV, RMVB, and M3U8.
- Video chat: Moderation (Video) accurately identifies and intercepts pornography, abuse, terrorism-related content, and advertisements in social and instant messaging scenarios.
  - Comprehensive detection: Moderation (Video) provides a multi-modal comprehensive review solution to parse the images, sounds, and text in videos.

  - Various types of video formats: Moderation (Video) supports video formats such as AVI, FLV, MP4, MPG, WMV, MOV, RMVB, and M3U8.
- Online education: Moderation (Video) accurately identifies and intercepts non-compliant content in online teaching, interaction, and recorded courses to protect the physical and mental health of users, especially minors.
  - Comprehensive detection: Moderation (Video) provides a multi-modal comprehensive review solution to parse the images, sounds, and text in videos.
  - Various types of video formats: Moderation (Video) supports video formats such as AVI, FLV, MP4, MPG, WMV, MOV, RMVB, and M3U8.

Audio Stream Moderation

- Live audio rooms

  Audio Stream Moderation is integrated into the Live audio platform to identify non-compliant content in live audio rooms in real time.

  Advantages:

  - Real-time: The content in live audio rooms can be monitored and analyzed in real time to ensure the order and security of the rooms.
  - Recognition of special sounds: Models for recognizing special sounds are available, such as asthma, moaning, and sensitive voiceprints.
- Social voice messages

  Audio Stream Moderation reviews voice messages sent by users on the social voice message platform in real time, identifies voice messages that contain malicious content in a timely manner, and helps you take action based on the review result, for example, deleting messages or forbidding users to speak.

  Advantages:

  - High accuracy: All scenarios are covered, preventing misoperations or omissions and defending against risks in real time.
  - Recognition of special sounds: Models for recognizing special sounds are available, such as asthma, moaning, and sensitive voiceprints.
- Online education

  Based on the education content and requirements, you can set appropriate review rules to help you identify audio streams that contain sensitive words or improper content, and detect and handle non-compliant content in a timely manner.

  Advantages:

  - High review efficiency: This service reduces manual review workloads, improves the accuracy of teaching content, and prevents incorrect or improper comments.
  - High accuracy: This service filters out inappropriate content and comments to ensure the security of teaching content.

4 Constraints

Text Moderation (V3)

- It is available in the **AP-Singapore** region.
- Text to be detected. The text is encoded using UTF-8. A maximum of 10,000 characters are allowed. If the text contains more than 10,000 characters, only the first 10,000 characters are detected.
- By default, the maximum number of concurrent API calls is 50 (a maximum of 50 requests within a second). To increase concurrency, **submit a service ticket**.
- There are fewer constraints at the Huawei Cloud International website than the Chinese Mainland website. For details, you can submit a service ticket to contact Huawei Cloud technical support.

Image Moderation (V3)

- It is available in the **AP-Singapore** region.
- It supports images in JPG, PNG, JPEG, WEBP, GIF, TIFF, TIF and HEIF formats.
- Each edge of an image must contain 20 to 6,000 pixels.
- If a Base64 string is transferred for an image, the encoded image cannot be larger than 10 MB (the original image cannot be larger than 7.5 MB). If a URL is transferred, the image cannot be larger than 10 MB.
- By default, the maximum number of concurrent API calls is 10 (a maximum of 10 requests within a second). To increase the concurrency, **submit a service ticket** to seek help from our professional engineers.
- There are fewer constraints at the Huawei Cloud International website than the Chinese Mainland website. For details, you can submit a service ticket to contact Huawei Cloud technical support.

Audio Moderation

- It is available in the **AP-Singapore** region.
- Audio files in WAV, MP3, AAC, AMR, 3GP, M4A, WMA, OGG, APE, FLAC, ALAC, WAVPACK and SILK_V3 formats are supported.
- The video size cannot exceed 200 MB.

- By default, the maximum number of concurrent API calls is 10 (a maximum of 10 requests within a second). To increase the concurrency, **submit a service ticket** to seek help from our professional engineers.

Video Moderation

- It is available in **AP-Singapore**.
- Formats such as AVI, FLV, MP4, MPG, WMV, MOV, WMA, RMVB and m3u8 are supported.
- The video file size cannot exceed 300 MB, and the video duration cannot exceed 2 hours.
- By default, the maximum number of concurrent API calls is 10 (a maximum of 10 requests within a second). To increase the concurrency, **submit a service ticket** to seek help from our professional engineers.
- There are fewer constraints at the Huawei Cloud International website than the Chinese Mainland website. For details, you can submit a service ticket to contact Huawei Cloud technical support.

Audio Stream Content Moderation

- It is available in the **AP-Singapore** region.
- Audio stream URL. Mainstream protocols such as RTMP, RTMPS, HLS, HTTP, and HTTPS are supported.
- By default, the maximum number of concurrent API calls is 10 (a maximum of 10 requests within a second). To increase the concurrency, **submit a service ticket** to seek help from our professional engineers.

5 Related Services

IAM

Identity and Access Management (IAM) provides Content Moderation with the user authentication and authorization function. For more information about IAM, see the *Identity and Access Management User Guide*.

OBS

Object Storage Service (OBS) is a stable, secure, efficient, and ease-of-use cloud storage service. Most Content Moderation APIs require data processing. You can use OBS to batch process data to improve data processing efficiency on the cloud.

Part of Content Moderation APIs can be temporarily authenticated or anonymously and publicly authorized to obtain data from OBS for processing. For more information about OBS, see the *Object Storage Service API Reference* and *Object Storage Service Developer Guide*.

6 Using Content Moderation

You can access Content Moderation on a web-based service management platform, that is, the management console, or using HTTPS-based APIs.

- You can subscribe to Content Moderation on the management console and view the number of successful API calls.
- If you access Content Moderation through open APIs, you need to integrate Content Moderation to a third-party system.

The procedure is as follows:

Step 1 Apply for a service.

You can apply for a service on the management console. For details about how to apply for a service, see **Applying for a Service** in the *Content Moderation API Reference*.

NOTE

- You only need to apply for a service once.
- This service is available only to enterprise users currently.

Step 2 Obtain request authentication.

You can use either of the following authentication methods when calling APIs:

- Token-based authentication: Requests are authenticated by using tokens. For details, see **Token-based Authentication** in the *Content Moderation API Reference*.
- AK/SK-based authentication: Requests are encrypted using the access key ID (AK) and secret access key (SK). AK/SK-based authentication is more secure. For details, see **AK/SK-based Authentication** in the *Content Moderation API Reference*.

Step 3 Call an API.

Content Moderation provides services through APIs. For details about how to call the APIs, see the **Content Moderation API Reference**.

Step 4 View service usage.

- You can view the number of API calls on the Content Moderation management console.

----End

7 Billing

Billing Items

Content Moderation is in commercial use. You can choose either pay-per-use billing or yearly/monthly packages. For details about Content Moderation pricing details, see **Product Pricing Details**.

Billing Modes

The pay-per-use and yearly/monthly billing modes are available.

- **Pay-per-use**

Content Moderation adopts tiered pricing based on the number of API calls. The tiered API calls are accumulated by calendar month. After a calendar month ends, the API calls are cleared. During the promotion period, each user can make API calls for different services free of charge each month. For details about the pricing, see **Product Pricing Details**.

NOTE

- This service is billed in pay-per-use mode by default. You can also purchase a discount resource package for a better price. However, if your usage exceeds the package quota, subsequently used resources will be billed on a pay-per-use basis.
- An API call is counted only when it is successfully called. Remaining free API calls at the end of the month do not roll over to subsequent months.
- Billing rule: tiered pricing based on the number of API calls (number of reviewed images). Each time an image is reviewed, a call is recorded. After a calendar month ends, the number of API calls is cleared and re-accumulated.
- Billing cycle: hourly. Bills are generally issued within 1 hour after each billing period ends, depending on how fast the system can process them.

- **Yearly/Monthly**

You can also purchase a discount resource package for a better price. However, if your usage exceeds the package quota, subsequently used resources will be billed on a pay-per-use basis. For details about the pricing, see **Content Moderation Pricing Details**. This billing mode provides a larger discount than pay-per-use and is recommended for long-term users.

Before making a purchase, learn and understand the following:

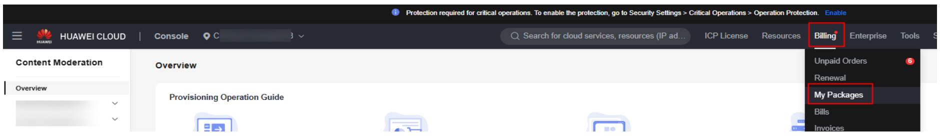
1. After you determine the required duration and the number of API calls, Content Moderation automatically calculates the fees you need to pay.
2. You can purchase and use multiple packages.

   For example, if you want to increase the number of API calls per month from 600,000 to 1.2 million, you can purchase a package that provides 600,000 API calls per month. For details about the package specifications, see **Content Moderation Pricing Details**.
3. Packages must be paid in full. They take effect immediately upon payment and become unavailable upon expiration. You cannot specify the date when a package takes effect.
4. A package is no longer available once it expires, even if the package has remaining number of API calls.

   For example, if you purchase a one-year package on January 1, the package automatically expires on January 1 in the next year. The validity period will not be extended and the fees cannot be refunded even if you do not make any API call within the validity period.
5. The fees for the API calls beyond the package quota are settled in pay-per-use mode according to tiered pricing.

NOTE

To view the remaining quota of your package, log in to the Content Moderation console and choose **Billing** > **My Packages** in the upper right corner.



Protection required for critical operations. To enable the protection, go to Security Settings > Critical Operations > Operation Protection. Enable

Overdue Payment

In pay-per-use mode, API fees are deducted every hour. If your account balance is insufficient to pay for the expense occurred in the last hour, your account will be in arrears.

If you top up your account within the retention period, the APIs will be available and billed from the original expiration date.

NOTE

If your account is in arrears, some operations will be restricted. You are advised to top up your account as soon as possible. The restricted operations are as follows:

- API calls purchased in pay-per-use mode cannot be used.
- Remaining API calls in a discount package can still be used, but the package cannot be subscribed again or renewed.
- Services cannot be subscribed.

Renewal

You can renew a resource package upon its expiration, or you can set auto-renewal rules for a resource package.

Expiration

- After a yearly/monthly package expires, you will be billed for subsequently used resources on a pay-per-use basis.
- If the account is not topped up or the resource package is not renewed before the retention period expires, your data will be deleted and cannot be recovered.