Tencent Cloud AI Digital Human

Introduction of Avatar

Product Documentation

Copyright Notice

Trademark Notice

Tencent Cloud

Service Statement

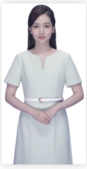
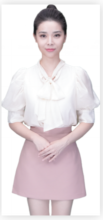
This document is intended to provide users with general information about Tencent Cloud's products and services only and does not form part of Tencent Cloud's terms and conditions. Tencent Cloud's products or services are subject to change. Specific products and services and the standards applicable to them are exclusively provided for in Tencent Cloud's applicable terms and conditions.

Contents

Introduction of Avatar
Introduction to Image Categories

Last updated：2025-03-13 17:06:20

Introduction to Image Categories

Avatar Type	Definition	Use Cases	Example
2D Premium	By recording motion materials in a professional studio, a digital human applied to broadcasting and interactive scenarios can be generated after about two weeks of training. The boutique image can randomly insert specified motions in the text, and the motions are diverse.	Applicable to customers in finance and media who have requirements for the image and motion of digital humans.	
2D small sample - general lip movement	A digital human is trained through a real-person video footage. The appearance of the digital human is consistent with that of the real person, and the mouth shape will use the general lip and teeth generated by the large model. The requirements for training video footage are lower. For details, see Image Recording Guide - General Mouth Shape.	Applicable to customers who have no requirements for the mouth shape of digital humans and no good shooting conditions.	
2D small sample - exclusive mouth shape	A digital human is trained through a real-person video footage. The appearance of the digital human is consistent with that of the real person, and the mouth shape will use the exclusive lip and teeth of the real person. The training video footage should have no other voices or obvious environmental sounds. For details, see Image Recording Guide - Exclusive Mouth Shape.	Applicable to customers who have requirements for the image replication of digital humans and good shooting conditions.	
2D small sample - high	A digital human is trained through a 4K real-person video footage. The material collection requirements and the final lip and	Applicable to large conferences, face-to-face dialogues,	

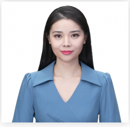
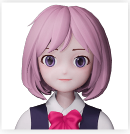
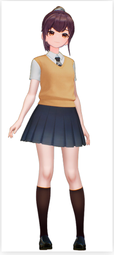
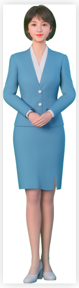
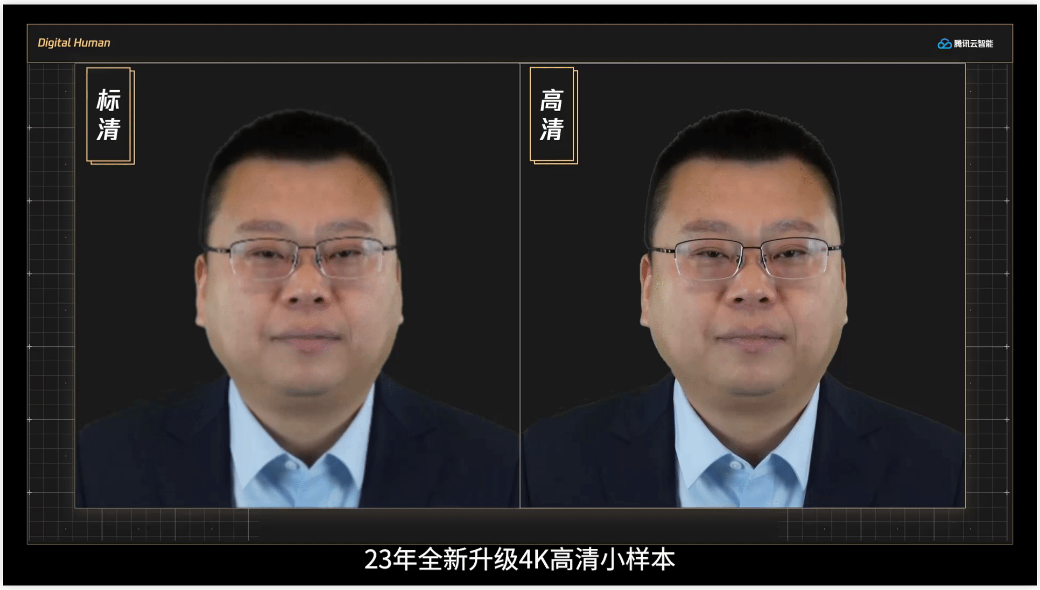
precision version	teeth effect are the same as those of 2D small sample (exclusive mouth shape). The resolution of the final digital human is upgraded to 4K. For details, see Image Recording Guide - High Precision Version.	product launch events, large screen scenarios.	
2D small sample - photo avatar	A digital human can be trained through a photo. This version is designed for low-cost and quick turnaround. Generally, it is ready for use within 10 minutes after material submission.	Applicable to pan-internet and entertainment scenarios.	
3D Cartoon	Set the facial features, hairstyle, clothing, accessories, etc. of the digital human according to the customer's requirements to complete the original painting. After the customer reviews and finalizes the final image, proceed with model making. After stages such as bone binding, rendering, and UE tuning, a digital human that covers interactive and broadcasting scenarios can be output.	Applicable to scenarios where there is an existing 2D mascot image and it is expected to be upgraded to a 3D image to provide services for users.	
3D Semi-Realistic	Set the facial features, hairstyle, clothing, accessories, etc. of the digital human according to the customer's requirements to complete the original painting. After the customer reviews and finalizes the final image, proceed with model making. After stages such as bone binding, rendering, and UE tuning, a digital human that covers interactive and broadcasting scenarios can be output.	It is suitable for scenarios requiring a certain realistic sense but not high precision, such as news reading and mobile smart customer service scenarios.	
3D Realistic	Set the facial features, hairstyle, clothing, accessories, etc. of the digital human according to the customer's requirements to complete the original painting. After the customer reviews and finalizes the final image, proceed with model making. After stages such as bone binding, rendering, and UE tuning, a digital human that covers interactive and broadcasting scenarios can be output.	It is suitable for scenarios requiring high realistic sense and high-precision display, such as brand promotion, large-screen interaction scenarios.	

Image Comparison

	2D Small Sample - General Lip Movement	2D Small Sample - Exclusive Mouth Shape	2D Small Sample - High Precision Version	2D Small Sample - Photo Avatar
Recording requirements	Record a video of at least 1 minute. There is no requirement for the sound of video shooting.	Record a video for at least 3 minutes. The recording environment needs to be quiet, and only the sound of the subject can be recorded.	The recording standard is the same as that of exclusive lip synchronization. The video resolution must be 4K.	Only one clear front photo of a person is required.
Delivery cycle	Deliver a demo within 1 day for customer effect confirmation. Customers can use it after clicking to confirm.	Deliver a demo within 2 days for customer effect confirmation. Customers can use it after clicking to confirm.	Deliver a demo within 3 days for customer confirmation. Customers can use it after clicking to confirm.	Available for use within 10 minutes.
Finished product effect	The general version uses lips and teeth generated by big data models.	The exclusive version records one's own lip movement, with better facial clarity.	Based on the effect of exclusive lip synchronization, output in 4K resolution for higher definition.	The photo avatar use lips and teeth generated by big dat models, and the body pose cannot sway slightly.
General lip movement vs Exclusive lip movement				

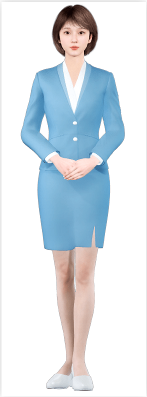
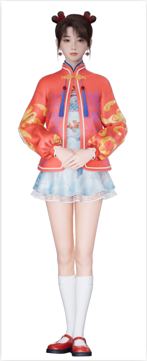
	通用口型 专属口型 前言 通用口型亮点 通用口型&专属口型对比 结束
General lip movement vs Photo avatar	 照片数字人只驱动整个头部，身体无法轻微晃动。
Exclusive lip movement vs High-precision version	

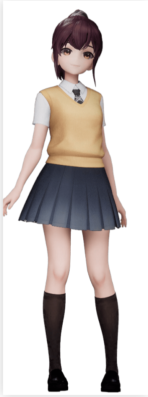
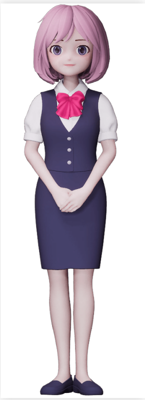
	 Digital Human 腾讯云智能 标清 高清 23年全新升级4K高清小样本

Basic Image Library

3D Basic Image Library

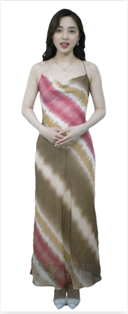
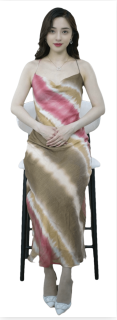
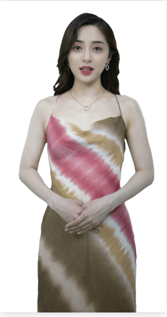
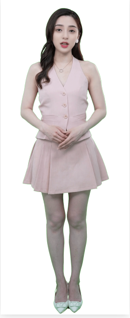
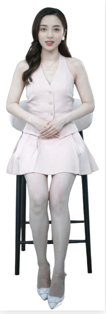
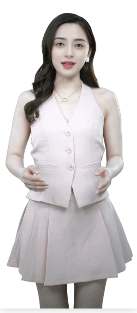
Last updated：2025-03-13 17:08:07

No.	Role	Avatar Type	Clothing	Avatar Example	With Action or Not	Supported Resolutions
1	yoyo	3D Realistic	Women's suit skirt		78 Motions	2560*1440 (Horizontal + Vertical) 1920*1080 (Horizontal + Vertical) 1280*720 (Horizontal + Vertical)
2	xiaowei	3D Realistic	Festival Clothing		18 Motions	2560*1440 (Horizontal + Vertical) 1920*1080 (Horizontal + Vertical) 1280*720 (Horizontal + Vertical)
3	aiyun	3D Semi-Realistic	College-style skirt suit		17 Motions	2560*1440 (Horizontal + Vertical) 1920*1080 (Horizontal + Vertical) 1280*720

						(Horizontal + Vertical)
4	dingdang	3D Cartoon	Customer Service Uniform		9 Motions	2560*1440 (Horizontal + Vertical) 1920*1080 (Horizontal + Vertical) 1280*720 (Horizontal + Vertical)

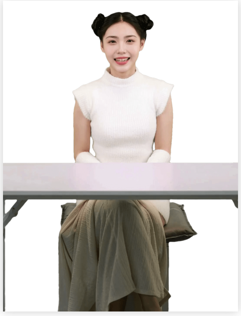
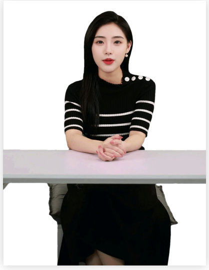
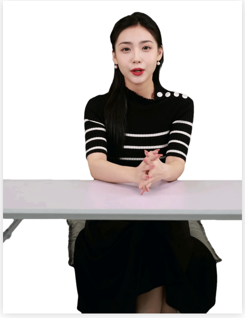
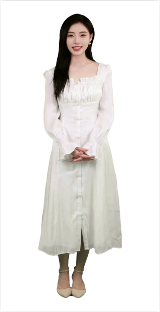
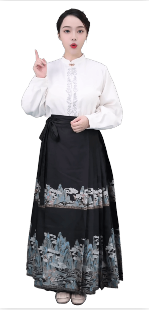
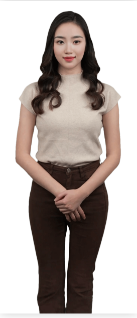
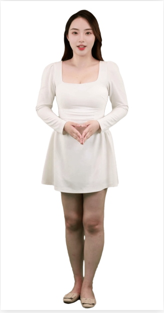
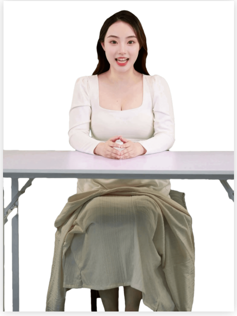
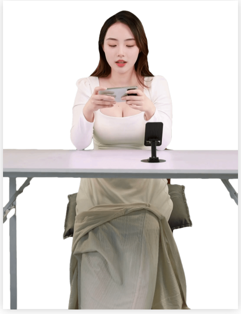
2D Small Sample (General Mouth Shape) Basic Image Library

Last updated：2025-03-21 09:52:42

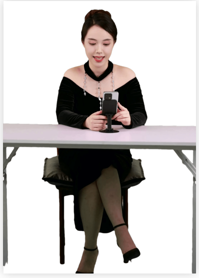
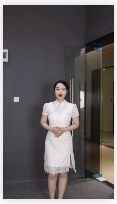
Yunxin - Highlights Long Dress - Full Body Standing Pose 	Yunxin - Highlights Long Dress - Seated Posture - Front 	Yunxin - Highlights Long Dress - Half-body Standing Pose - Front 
Yunxin - Pink Dress Set - Full Body Standing Pose 	Yunxin - Pink Dress Set - Seated Posture - Front 	Yunxin - Pink Dress Set - Half-body Standing Pose - Front 
Yunxin - Yellow Denim Suit - Full Body Standing Pose	Yunxin - Yellow Denim Suit - Seated Posture - Front	Yunxin - Yellow Denim Suit - Half-bod Standing Posture - Front

Yunyou - Colorful Suit - Full Body Standing Pose 	Yunyou - Colorful Suit - Seated Posture - Front 	Yunyou - Colorful Suit - Half-body Standing Posture - Front
Yunyou - Collegiate Suit - Full Body Standing Pose 	Yunyou - Collegiate Suit - Seated Posture - Front 	Yunyou - Collegiate Suit - Half - body Standing Posture - Front
Yunhan - Yoga Suit - Full Body Standing Pose	Yunhan - Yoga Suit - Seated Posture - Front	Yunhan - Yoga Suit - Half-body Standing Posture - Front

Yunhan - Denim Skirt Suit - Full Body Standing Pose 	Yunhan - Denim Skirt Suit - Seated Posture - Front 	Yunhan - Denim Skirt Suit - Half Body Standing Pose - Front
Yunxi - Blue Suit - Full Body Standing Pose 	Yunxi - Blue Suit - Standing Pose - Side 	Yunxi - Blue Suit - Seated Posture - Front
Zhiqi - White T-shirt - Seated Posture	Zhiqi - Striped Black Clothes - Seated Posture with Loose Hair	Zhiqi - Striped Black Clothes - Seated Posture

		
Zhiqi - White Long Dress - Standing Posture 	Yunshu - Horse-face Skirt - Full Body Standing Pose 	Yunxi - Casual Pants - Standing Posture 
Zhixuan - White Dress - Standing Posture 	Zhixuan - White Dress - Seated Posture 	Zhixuan - White Dress - Seated Posture on Mobile Phone in Landscape Mode 
Tengfeng - Gray Suit - Standing Posture	Tengfeng - Black Shirt - Standing Posture	Tengfeng - Suit - Standing Posture

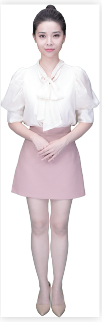
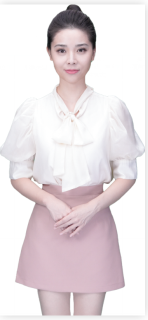
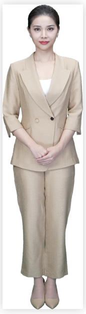
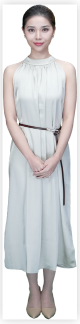
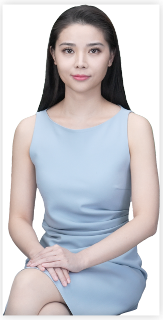
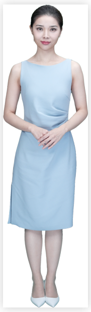
Tengfeng - Suit - Seated Posture Browsing Mobile Phone 	Yunfan - Suit Skirt - Half Body Standing Pose 	Yunfan - Suit Skirt - Horizontal Screer Half Body
Yunni - Qipao - Full Body Standing Pose 	Yunshu - Long Dress - Full Body Standing Pose 	Yunman - Qipao - Standing Pose
Zhixuan - Plaid Skirt - Full Body Standing Pose	Zhixuan - Plaid Skirt - Seated Posture	Zhixuan - Flower-print Skirt - Seated Posture

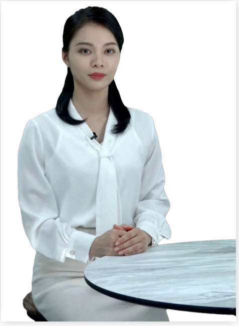
Zhixuan - Black Dress - Standing Posture 	Zhixuan - Black Dress - Seated Posture on Mobile Phone in Portrait Mode 	Zhixuan - Black Traditional Chinese Clothing - Full Body Standing Posture
Financial Xiao Teng 	Auto Xiao Teng - Suitable for Broadcast Video within Two Minutes 	Yunjing - Qipao - Seated Posture Interview - Front
Yunni - Qipao - On-site 	Auto Xiao Teng 	Yunshu - Horse-face Skirt - On-site
Yunjing - Outdoor Interview - Suitable for broadcast video within	Yunjing - Outdoor Interview - Suitable for broadcast video	Yunjing - Outdoor Interview - Suitable for broadcast video within 1.5 minutes

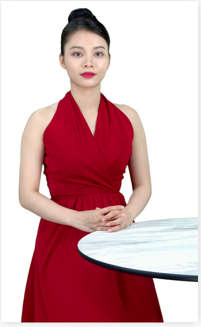
30 seconds 	within 1 minute 	

2D Small Sample (Exclusive Mouth Shape) Basic Image Library

Last updated：2025-03-21 10:27:20

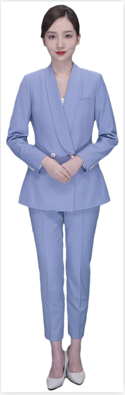
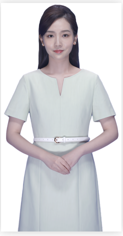
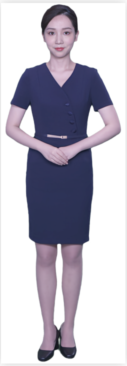
weilan - pink mini skirt suit - full body standing pose 	weilan - pink mini skirt suit - half-body standing posture 	weilan - champagne suit - full body standing pose 	weila postu
weilan - off-the-shoulder dress - full body standing pose 	weilan - morandi blue sleeveless dress - seated posture 	weilan - morandi blue sleeveless dress - full body standing pose 	weila skirt :
Xiaozhe - black suit - full body standing pose	Jing'er - black dress - seated posture	Jing'er - black dress - half-body standing posture	Jing'e body

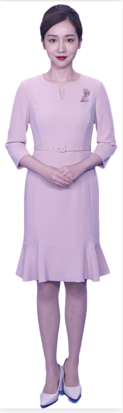
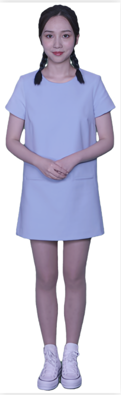
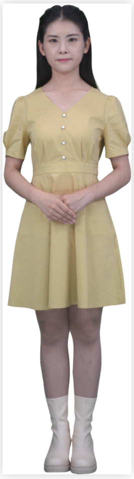
Jing'er - gray pleated skirt - full body standing pose 	Jing'er - gray pleated skirt - seated posture - high stool 	Jing'er - gray-black casual dress - full body standing posture 	Jing'e half-b
Manwen - white shirt - full body standing pose 	Manwen - white shirt - seated on a high stool 	Manwen - white shirt - seated beside the table 	Manv front
Manwen - white lab coat - full body standing pose	Manwen - white lab coat - seated on a high stool	Manwen - white lab coat - seated beside a table	Manv front

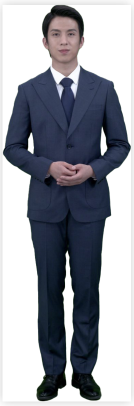
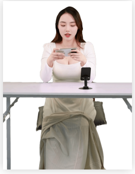
Manwen - white suit - full-body standing posture 	Manwen - white suit - seated on a high stool 	Manwen - white suit - seated beside the table 	Manv table
Manwen - red dress - full body standing pose 	Manwen - red dress - seated on a high stool 	Manwen - red dress - seated beside a table 	Manv of a t
Xiaozhe - light khaki shirt - seated on a high stool	Xiaozhe - light khaki shirt - half-body standing posture		

2D Boutique Basic Image Library

Last updated：2025-03-21 10:34:53

No.	Role	Clothing	Posture	Image Example	With or Without Motion	Supported Voice Types	Supported Resolution
1	Yunxuan	Blue suit	Full body standing posture		8 Actions	19 Voice types	1920*1080 (Horizontal 1280*720 (Horizontal 1080*1920 (Vertical) 540*960 (Vertical)
2	Yunxuan	Green skirt	Half-body standing posture		8 Actions	19 Voice types	1920*1080 (Horizontal 1280*720 (Horizontal 1080*1920 (Vertical) 540*960 (Vertical)
3	Yunxuan	Customer service uniform	Full body standing posture		12 Actions	19 Voice types	1920*1080 (Horizontal 1280*720 (Horizontal 1080*1920 (Vertical) 540*960 (Vertical)
4	Yunxuan	Pink long	Full		11	19 Voice	1920*1080

		dress	body standing posture		Actions	types	(Horizontal 1280*720 (Horizontal 1080*1920 (Vertical) 540*960 (Vertical)
5	Yunxuan	Blue T-shirt dress	Full body standing posture		8 Actions	19 Voice types	1920*1080 (Horizontal 1280*720 (Horizontal 1080*1920 (Vertical) 540*960 (Vertical)
6	Shu Yu	Yellow skirt	Full body standing posture		8 Actions	19 Voice types	1920*1080 (Horizontal 1280*720 (Horizontal 1080*1920 (Vertical) 540*960 (Vertical)
7	Shu Yu	Blue skirt	Full body standing posture		8 Actions	19 Voice types	1920*1080 (Horizontal 1280*720 (Horizontal

							1080*1920 (Vertical) 540*960 (Vertical)
8	Tengyu	Blue suit	Full body standing posture		8 Actions	19 Voice types	1920*1080 (Horizontal 1280*720 (Horizontal 1080*1920 (Vertical) 540*960 (Vertical)
9	Tengyu	Blue suit	Seated posture		No action	19 Voice types	1920*1080 (Horizontal 1280*720 (Horizontal 1080*1920 (Vertical) 540*960 (Vertical)
10	YunWu	Red suit	Half-body standing posture		8 Actions	19 Voice types	1920*1080 (Horizontal 1280*720 (Horizontal 1080*1920 (Vertical) 540*960 (Vertical)

Guide on Avatar and Voice Clone
Overview

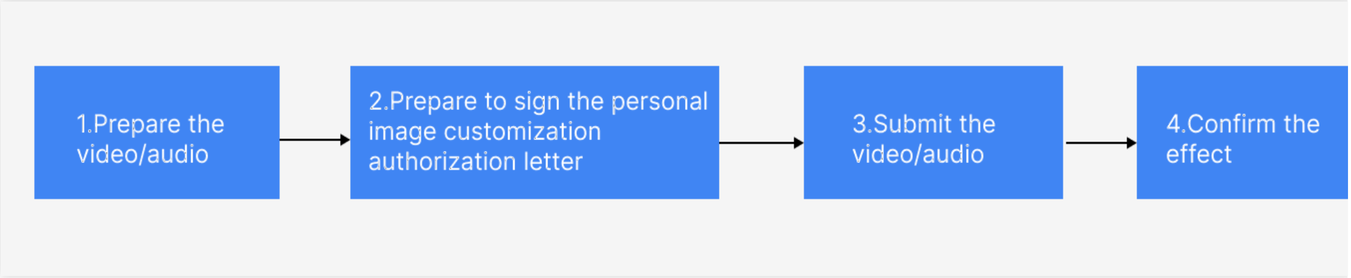
Last updated：2024-09-18 20:42:04

Overview

Provide a 3-minute video to quickly generate an Avatar identical to the real person, with facial features, movements, and expressions fully mimicking the original. You only need to input text or audio to produce an Avatar broadcasting video quickly.

Note:
The final customized Avatar will closely match the makeup, skin tone, facial expressions, and movements of the person in the provided video. The original lighting from the video will also be retained. If you require beautification or brightness adjustments, you can enhance the original video before uploading it. Ensure that the video is edited to your satisfaction before submission.

Process Guide


1.Prepare the video/audio
2.Prepare to sign the personal image customization authorization letter
3.Submit the video/audio
4.Confirm the effect

1. Prepare video/audio material.

Avatar customization: Quickly generate an Avatar identical to the real person, with facial features, movements, and expressions fully mimicking the original.

	Method 1: Record the Material	**Method 2: Use the Existing Material**
Restricted	Record a 3-5 minute video following the guide on Avatar customization. For recording and post-production, see Key Shooting Points and Post-production Guide.	Use an existing video with spoken content of at least 1 minute (unedited version).
Note	Video quality: The face should be clear and not blurry, with well-defined edges even when zoomed in. The video should be stable with no shaking. Model performance: The eyes should be looking directly at the camera, with no significant head turns or tilts, and the face should remain unobstructed throughout the video. Filming key points: The video should begin with 1-3 seconds of silence with the mouth closed. The entire video must be unedited, with no frame skips, and the total length must exceed 3 minutes. Environmental voice: No other voices should be mixed in, and there should be no significant background noise. The audio and video must be synchronized, with the voice matching the lip movements.	This method does not support background replacement; the background will remain consistent with that of the original video.

Voice clone: Quickly generate a timbre that matches the original speaker's voice.

Record 100 sentences of audio (usually 15-20 minutes), recommended format is wav for lossless compression.

You can see VRS Recording Guide.

Note:

Voice clone is currently not available as a standalone service; it must be paired with Avatar customization for a complete solution.

2. Prepare a personal Avatar customization authorization letter.

To protect the legal rights of customers and the individual being customized, according to the Provisions on the Administration of Deep Synthesis of Internet-Based Information Services, appropriate measures need to be taken to ensure that legal authorization and consent from the user have been obtained. After the user authorization is obtained, the customer can provide the authorization materials to Tencent Cloud through online APIs or offline methods. Each Avatar requires the submission of the individual's authorization materials, and two authorization modes are supported: video-verbal authorization and written authorization **(preferably both authorization modes should be provided).** Click to download the authorization file template: Image Authorization Letter.

3. Submit video/audio material files.

Before submission, it is necessary to perform a self-check on the material. Once the custom material passes the self-check, it can be submitted for training through the API interface or the Digital Human platform.
For details, see Custom Material Submission Guidelines.

4. Confirm the customization effect and service usage.

You can check the customization progress of the Avatar through Avatar Image Customization and Voice Clone API Documentation provided in the API documentation. Service usage can be accessed through the broadcasting or live streaming aPaaS APIs.

Avatar Recording Guide - Studio Avatar

Last updated：2025-04-09 14:58:43

I. Custom Material Self-Check Items

For Avatar customization, you need to submit a 3-5 minute real-person video with spoken content. Before submission, ensure you check each of the following self-check items:

1. Video quality: The face should be clear and not blurry, with well-defined edges even when zoomed in. The video should be stable with no shaking.
2. Model performance: The eyes should be looking directly at the camera, with no significant head turns or tilts, and the face should remain unobstructed throughout the video.
3. Filming key points: The video should begin with 1-3 seconds of silence with the mouth closed. The entire video must be unedited, with no frame skips, and the total length must exceed 3 minutes.
4. Environmental voice: No other voices should be mixed in, and there should be no significant background noise. The audio and video must be synchronized (the voice should match the lip movements).
5. Filming background: If the cutout is required, a green screen or white screen background must be provided. The green or white screen should completely cover the background and be free of any other objects.

Video Format Requirements:

1. The video size must not exceed 5 GB, with a duration of no less than 3 minutes and no more than 10 minutes.
2. The video format should be either MP4 or MOV.
3. The video resolution should be 1080P or 4K (3840 x 2160) with an aspect ratio of 16:9 (or 9:16), as shown below.

Model sample

16:9	9:16

4. The video frame rate should be no less than 25 fps and no more than 60 fps.
5. The person's head in the video must be upright. If the person is positioned horizontally, the video should be rotated to correct the orientation.

Video guide: Filming guide

II. Filming Guide

Filming Location Setup

1. Location selection

Notes:

If there is a need for background replacement in post-production, use a green screen or white screen for filming. If a fixed background is desired, choose an appropriate environment for on-site filming; the background will be retained in the videos generated subsequently.

On-site filming: Choose a well-lit, quiet **on-site** room with no background noise for recording. (On-site filming means the background is fixed and cannot be replaced in post-production. For outdoor filming, use a microphone to ensure clear voice without noise.)

Green screen filming: Choose a well-lit, quiet **green screen** room with no background noise for recording.

White screen filming: Choose a well-lit, quiet room with **a white wall or white screen** for recording. (White screen filming currently does not support shooting with tables or chairs.)

2. Model clothing and style selection

Model: The model should have well-defined facial features, be attractive, possess a good presence, have clear speech, and act naturally. Preference is given to models with extensive on-camera experience.

Clothing:

On-site filming: No specific color requirements for clothing.

Green screen filming: Avoid reflective materials or clothing with checkered or striped patterns. Do not wear clothing in colors similar to green (such as yellow, green, or yellow-green) to prevent issues with cutting out the background.

White screen filming: Avoid wearing white or similar-colored clothing. White clothing is acceptable if it is not on the body’s edges (e.g., an inner layer under a suit).

Hairstyle: The hairstyle should be neat, avoiding noticeable partings and stray hairs. Avoid wearing dangling earrings. (This requirement applies only to the material for green screen and white screen filming; there are no such restrictions for on-site filming.)

On-site filming example:

Green screen filming example:

White screen filming example:

Dual Avatar example: (Currently in testing, stay tuned for updates)

3. Pre-Filming Model Preparations

3.1 Determine the posture

Select the posture of the model. Pay attention to the position and proportion of the figures in the figure.

Front sitting: Special attention should be paid to leaving sufficient space for hand exercises.

Side sitting: Special attention should be paid to fully revealing the corners of the mouth.

Keep a front half-body posture when standing.

Completely front posture

3.2 Confirm the clothing, style, and model's location (Special attention to green screen shooting)

a) Avoid wearing green or similar color clothing, accessories (including any green or reflective accessories)

b) Avoid messy and hairy hairstyles. (Impact on the effect of the green screen backdrop and subsequent use)

c) Avoid wearing clothes with fine mesh, stripes, transparent tulle, tattered skirts, etc., to avoid situations such as transparent or moiré patterns during shooting.

d) Determine in advance whether to wear glasses or contact lenses (to avoid problems with prompters and glasses that reflect light).

e) You can wear makeup (avoid using flash and highlights).

f) Avoid using swinging earrings, but you can wear them if they are completely covered by hair and not exposed against a green screen background.

Recommend the model to stand 1.5 to 2 meters away from the green screen to avoid green light reflection from the body edge.

3.3 Precautions for models

Keep your eyes on the lens (unless you are shooting a side view of a digital human).

When you receive the signal from the photographer to start shooting, stand naturally, keep silent, blink and nod slightly, and maintain this silence for more than 3 seconds.

Start speaking in a natural state, maintain natural head and hand movements, and continue for more than 3 minutes. Even if you make a mistake, you don't need to stop. Just continue speaking.

A prompter can be used, but it is more advisable not to look at it; text can be used for introduction, storytelling, etc.

Simulate the state of natural speaking with natural head and hand movements without speaking, and maintain for more than 1 minute (Items 3 and 4)

The interactive digital human should try to speak for 6 to 30 seconds, then perform a common gesture. This gesture should be completed within 2 to 3 seconds. Afterward, the digital human should speak again for 6 to 30 seconds, and then perform another gesture. This process should be repeated until the video recording is completed. In the last part of the presentation, gestures are not required.

When doing head movements, note to keep your mouth wide open throughout the process (for example, do not turn your head to one side). Do not make large hand movements or gestures with clear indications (for example, hello, goodbye, digits 1, 2, 3, etc.).

3.4 Other Recommendations

If conditions permit, you can shoot multiple postures, from multiple angles, and in multiple sets of clothing materials for selection later. It should be noted that uploading multiple materials requires the use of training quotas for multiple models.

This material can be used in conjunction with appropriate facial beautification operations, but it is recommended to avoid the facial slimming feature (this may lead to changes in face size).

In principle, voice cloning requires more than 10 seconds. The longer the time, the higher the recovery degree. Use a natural speaking tone and avoid using the tone when reading. Speak loudly (unless you want to train the model in this way).

During the shooting process, the nodding posture should maintain the minimum range and reduced frequency to ensure the delicacy and naturalness of the picture.

4. Filming Equipment and Lighting

Ensure that the camera remains stable without shaking during filming and that the lighting does not change significantly during the recording process.

Camera shooting settings: Resolution 4K or 1080P, frame rate 30fps, normal exposure; Mobile phone shooting settings: Video mode (non-cinematic effect), resolution 4K or 1080P, frame rate 30fps, disable PAL format, disable HDR mode

The green screen should be smooth, without wrinkles, and should fully cover the entire frame.

Shoot in one continuous take, without splicing video; no obstruction by the camera; when you detect abnormal scenes (for example, people moving out of the frame, etc.), interrupt and reshoot promptly.

5. Mobile Filming Standards

The preferred device for mobile filming is an iPhone. The specific parameters for filming are as follows: use the rear camera in video mode (not cinematic mode), set the zoom to 1x, resolution to 4K, and frame rate to 30 fps. Disable PAL format, HDR mode, and auto FPS. These settings can be adjusted under Settings > Camera > Record Video. The specific settings are shown in the figure below:

Filming and Recording

1. Video recording position

2. Real-time monitoring and preview during filming

You can use software like OBS for real-time preview of the cutout effect. This allows you to detect issues such as reflective accessories or green spills on the face and clothing in advance. Adjustments can be made on set in real-time to avoid repeated recordings and delays in the customization process.

3. Filming and recording (audio recording required simultaneously)

Shot selection: If the final video will be used in a vertical format, it is recommended to shoot in the vertical mode; the same applies to horizontal format. When the entire body is filmed, ensure that the subject appears as large as possible, and keep hand movements within the frame.

Recording process:

1. After the recording is started, the model should keep their mouth closed for 1-3 seconds, maintaining a still posture.
2. Next, the model should speak naturally for 3-5 minutes, avoiding repetition of the same script. During the speech, the model can make subtle, natural movements, keeping their eyes focused on the camera without looking sideways.
3. After the speech is finished, stop the recording.

Movement suggestions: While speaking, the model can make neutral, versatile hand gestures. If unsure about gestures, the model can cross their hands in front of them. Ensure that the gestures are small, slow, and gentle, avoiding any obstruction of the neck or face. Gestures should be non-specific and adaptable to all types of text. (If the Avatar is ultimately used for real-time interaction scenes, there are additional requirements for hand movements. See the fifth section of this page for details.)

III. Post-Processing

1. Editing

Trim the beginning and end of the video to remove any unnecessary footage.

Ensure that the frame rate of the editing project matches that of the recorded material to avoid misalignment between the audio and lip movements.

2. Color correction and beautification

Correct any imperfections in the footage to ensure the model looks their best, but retain the natural texture of the model's skin. Avoid making the skin appear too white or too smooth.

3. Audio adjustment

If the audio in the video contains noise, it needs to be removed to ensure good voice quality. The synchronized audio should be clear.

4. Cutout

If you have cutout capabilities, you can perform the cutout process on the original video in advance. The video output options will vary based on the provided video material.

Case 1: Provide a green background video that has already undergone cutout (Video 2 in the figure below). The Avatar side will directly output the video with the green background (Video 3 in the figure below).

Clients can provide a green background video that has undergone the cutout for training. The Avatar side will directly use the green background as the final output video background. This approach offers higher customization efficiency and shorter delivery times. The cutout guide is as follows:

Remove the green screen background and eliminate any green reflections on the actor. Check the video against other background colors to ensure a clean cutout, making sure it can seamlessly adapt to any background.

After the clean cutout, fill the background with a pure green color, #00ff00 (R:0, G:255, B:0).

In the Avatar interaction & broadcasting API, the output video and video stream do not support background replacement, meaning: (1) Background replacement is not supported in the output. (2) Transparent background webm videos are not supported. After receiving the output video from the Avatar, the client will need to perform further green screen removal in their use cases.

Case 2: In addition to providing the original recorded video, an additional alpha channel video is provided (Video 2 in the figure below). On the Avatar side, background replacement in the output is supported (Video 3 in the figure below).

You need to provide both Video 1: Original Recorded Video (which can also be a processed video) and Video 2: Alpha Channel Video. The resolution and duration of these two videos should be exactly the same.

In this case, the videos and video streams output by the Avatar interaction & broadcasting API support background replacement.

IV. Recording Requirements for Avatars in Interactive Scenes

If the Avatar is intended for real-time interactive scenes, there are additional requirements for the model's hand movements during the 3-5 minute video recording. The specific requirements are as follows:

Each movement should be brief (see Hand Movement Illustration). **After the movement is completed, quickly return hands to the starting position** (see Hand Position Reset Illustration). **There should be at least 5 seconds between movements.**

Note: The final Avatar will replicate the movements exactly as they were performed during filming. If no movements are performed throughout the recording, the final Avatar will also have no hand movements.

1. **Hand movement illustration:**

The model's hands can perform some general movements. **After the movement is completed, quickly return the hands to the starting position, with each movement lasting no more than 2 seconds.** This segment will be used for speaking mode in the Avatar interactive scenes. The illustration is as follows:

2. **Hand position reset illustration:**

In this segment, while the model continues speaking naturally, the hands should avoid making any noticeable movements. This segment will be used for the listening/waiting mode in the Avatar interactive scenes. The illustration is as follows:

3. The reference video for recording the demo is as follows:

Avatar Recording Guide - Instant Avatar

Last updated：2025-04-09 15:05:33

I. Custom Material Self-Check Items

For Avatar customization, you need to submit a real-person video of at least 1 minute in length. Before submission, ensure that you check each of the following self-check items:

1. Video quality: The face should be clear and not blurry, with well-defined edges even when zoomed in. The video should be stable with no shaking.

2. Model performance: The eyes should be looking directly at the camera, with no significant head turns or tilts, and the face should remain unobstructed throughout the video.

3. Filming key points: The video should begin with 1-3 seconds of silence with the mouth closed. The entire video must be unedited, with no frame skips, and the total length must exceed 1 minute.

4. Environmental voice: No audio recording is required. The model can simply keep their mouths naturally closed throughout the video.

5. Filming background: If the cutout is required, a green screen or white screen background must be provided. The green or white screen should completely cover the background and be free of any other objects.

Video Format Requirements:

1. The video size should not exceed 5 GB, with a duration of no less than 1 minute and no more than 10 minutes.

2. The video format should be either MP4 or MOV.

3. The video resolution should be 1080P or 4K (3840 x 2160), with an aspect ratio of 16:9 (or 9:16), as shown below.

Model example

16:9	9:16

4. The video frame rate should be no less than 25 fps and no more than 60 fps.

5. The person's head in the video must be upright. If the person is positioned horizontally, the video should be rotated to correct the orientation.

Video Guide: Filming Guide

II. Filming Guide (Text Version)

Filming Location Setup

1. Location selection

Notes:

If there is a need for background replacement in post-production, use a green screen or white screen for filming. If a fixed background is desired, choose an appropriate environment for on-site filming; the background will be retained in the videos generated subsequently.

On-site filming: Choose a well-lit, stable **indoor/outdoor location** for recording. **No specific audio requirements are necessary.** (On-site filming means the background is fixed and cannot be replaced with another background in post-production.)

Green screen filming: Record the material in front of a well-lit, stable **green screen**. **No specific audio requirements are necessary.**

White screen filming: Record in a well-lit, stable, quiet room with a white wall or white screen. **No specific audio requirements are necessary**. (White screen filming currently does not support shooting with tables or chairs.)

2. Model clothing and style selection

Model: The model should have well-defined facial features, be attractive, possess a good presence, have clear speech, and act naturally. Preference is given to models with extensive on-camera experience.

Clothing:

On-site filming: No specific color requirements for clothing.

Green screen filming: Avoid reflective materials or clothing with checkered or striped patterns. Do not wear clothing in colors similar to green (such as yellow, green, or yellow-green) to prevent issues with cutting out the background.

White screen filming: Avoid wearing white or similar-colored clothing. White clothing is acceptable if it is not on the body's edges (e.g., an inner layer under a suit).

Hairstyle: The hairstyle should be neat, avoiding noticeable partings and stray hairs. Avoid wearing dangling earrings. (This requirement applies only to the material for green screen and white screen filming; there are no such restrictions for on-site filming.)

On-site filming example 1:

On-site filming example 2:

Green screen filming example:

White screen filming example:

Dual Avatar example: (Currently in testing, stay tuned for updates)

3. Pre-Shooting Model Preparations

3.1 Determine the posture

Select the posture of the model. Pay attention to the position and proportion of the figures in the figure.

Front sitting: Special attention should be paid to leaving sufficient space for hand exercises.

Side sitting: Special attention should be paid to fully revealing the corners of the mouth.

Keep a front half-body posture when standing.

Completely front posture

3.2 Determine the clothing, style, and model's position (Special attention to green screen shooting)

a) Avoid wearing green or similar color clothing, accessories (including any green or reflective accessories)

b) Avoid messy and hairy hairstyles. (Impact on the effect of the green screen backdrop and subsequent use)

c) Avoid wearing clothes with fine meshes, stripes, transparent tulle, tattered skirts, etc., to avoid situations such as transparent and moiré patterns in shooting.

d) Determine in advance whether to wear glasses or contact lenses (to avoid problems with prompters and glasses that reflect light).

e) You can wear makeup (avoid using flash and highlights).

f) Avoid using swinging earrings, but you can wear them if they are completely covered by hair and not exposed against a green screen background.

g) It is recommended that the model stand 1.5 to 2 meters away from the green screen to avoid green light reflections from the body edges.

3.3 Precautions for models

Keep your eyes on the lens (unless you are shooting a side view of a digital human).

When you receive the signal from the photographer to start shooting, stand naturally, keep silent, blink and nod slightly, and maintain this silent state for more than 3 seconds.

Without speaking, use natural head and hand movements to simulate speaking naturally and maintain this for more than 1 minute (items 3 and 4).

The interactive digital human should try to speak for 6 to 30 seconds, then perform a common gesture. This gesture should be completed within 2 to 3 seconds. Afterward, the digital human should speak again for 6 to 30 seconds, and then perform another gesture. This process should be repeated until the video recording is completed. No gestures are required in the last part of the presentation.

When doing head movements, note to keep your mouth wide open throughout the process (for example, do not turn your head to the side). Do not make large hand movements or gestures with clear indications (for example, hello, goodbye, numbers 1, 2, 3, etc.).

3.4 Other Recommendations

If conditions permit, you can shoot multiple postures, from multiple angles, and in multiple sets of clothing materials for selection later. It should be noted that uploading multiple materials requires the use of training quotas for multiple models.

This material can be used in conjunction with appropriate facial beautification operations, but it is recommended to avoid the facial slimming feature (this may lead to changes in face size).

In principle, speech cloning requires more than 10 seconds. The longer the time, the higher the recovery degree. Use a natural speaking tone and avoid using the tone when reading. Speak loudly (unless you want to train the model in this way).

During the shooting process, the nodding posture should maintain the minimum range and reduced frequency to ensure the delicacy and naturalness of the picture.

4. Filming Equipment and Lighting

Ensure that the camera remains stable without shaking during filming and that the lighting does not change significantly during the recording process.

The recording resolution should be 1080p or higher, and HDR mode should not be enabled during filming.

The green screen should be smooth, without wrinkles, and should fully cover the entire frame.

5. Mobile Filming Standards

The preferred device for mobile filming is an iPhone. The specific parameters for filming are as follows: use the rear camera in video mode (not cinematic mode), set the zoom to 1x, resolution to 4K, and frame rate to 30 fps. Disable PAL format, HDR mode, and auto FPS. These settings can be adjusted under Settings > Camera > Record Video. The specific settings are shown in the figure below:

Filming and Recording

1. Video recording position

2. Real-time monitoring and preview during filming

You can use software like OBS for real-time preview of the cutout effect. This allows you to detect issues such as reflective accessories or green spills on the face and clothing in advance. Adjustments can be made on set in real-time to avoid repeated recordings and delays in the customization process.

3. Filming and recording (no audio required)

Shot selection: If the final video will be used in a vertical screen format, it is recommended to shoot in the vertical mode; the same applies to horizontal format. When the entire body is filmed, ensure that the subject appears as large as possible within the frame.

Recording process: Choose any of the following options for recording. During the process, ensure that the eyes do not look sideways. Keep a direct gaze at the camera, and make sure movements stay within the frame.

Option 1: At the start of the recording, the model should keep their mouths closed and perform natural, subtle movements.

Option 2: At the start of the recording, the model should first keep their mouth closed for 1-3 seconds. After that, the model can speak naturally (ensuring that the lip movements are not too exaggerated) while performing natural, subtle movements.

You may stop recording once the duration exceeds 1 minute.

Movement suggestions: While speaking, the model can make neutral and versatile hand gestures. If unsure about gestures, the model can cross their hands in front of them. Ensure that the gestures are small, slow, and gentle, without obstructing the neck or face. Avoid gestures that have specific meanings or directional intent, as they need to be suitable for all types of text. (If the Avatar is ultimately used for real-time interaction scenarios, there are additional requirements for hand movements. See the fourth section of this page for details.)

III. Post-Processing

1. Editing

Trim the beginning and end of the video to remove any unnecessary footage.

Ensure that the frame rate of the editing project matches that of the recorded material to avoid misalignment between the audio and lip movements.

2. Color correction and beautification

Correct any imperfections in the footage to ensure the model looks their best, but retain the natural texture of the model's skin. Avoid making the skin appear too white or too smooth.

3. Audio adjustment

If the audio in the video contains noise, it needs to be removed to ensure good voice quality. The synchronized audio should be clear.

4. Cutout

If you have cutout capabilities, you can perform the cutout process on the original video in advance. The video output options will vary based on the provided video material.

Case 1: Provide a green background video that has already undergone cutout (Video 2 in the figure below). The Avatar side will directly output the video with the green background (Video 3 in the figure below).

Clients can provide a green background video that has undergone the cutout for training. The Avatar side will directly use the green background as the final output video background. This approach offers higher customization efficiency and shorter delivery times. The cutout guide is as follows:

Remove the green screen background and eliminate any green reflections on the actor. Check the video against other background colors to ensure a clean cutout, making sure it can seamlessly adapt to any background.

After the clean cutout, fill the background with a pure green color, #00ff00 (R:0, G:255, B:0).

In the Avatar interaction & broadcasting API, the output video & video stream do not support background replacement, meaning: (1) background replacement is not supported in the output; (2) transparent background webm videos are not supported. After receiving the output video from the Avatar, the client needs to perform further green screen removal in their use cases.

Case 2: In addition to providing the original recorded video, an additional video with an alpha channel (Video 2 in the figure below) is provided. On the Avatar side, background replacement in the output is supported (Video 3 in the figure below)

You need to provide both Video 1: Original Recorded Video (which can also be a processed video) and Video 2: Alpha Channel Video. The resolution and duration of these two videos must be exactly the same.

In this case, the videos and video streams output by the Avatar interaction & broadcasting API support background replacement.

IV. Recording Requirements for Avatars in Interactive Scenes

If the Avatar is intended for real-time interactive scenes, there are additional requirements for the model's hand movements during the 3-5 minute video recording. The specific requirements are as follows:

Each movement should be brief (see Hand Movement Illustration). **After the movement is completed, quickly return hands to the starting position** (see Hand Position Reset Illustration). **There should be at least 5 seconds between movements.**

Note: The final Avatar will replicate the movements exactly as they were performed during filming. If no movements are performed throughout the recording, the final Avatar will also have no hand movements.

1. **Hand movement illustration:**

The model's hands can perform some general movements. **After the movement is completed, quickly return the hands to the starting position, with each movement lasting no more than 2 seconds.** This segment will be used for speaking mode in the Avatar interactive scenes. The illustration is as follows:

2. **Hand position reset illustration:**

In this segment, while the model continues speaking naturally, the hands should avoid making any noticeable movements. This segment will be used for the listening/waiting mode in the Avatar interactive scenes. The illustration is as follows:

3. The reference video for recording the demo is as follows:

Avatar Recording Guide - 4K Version

Last updated：2024-09-18 20:42:21

The overall recording requirements are the same as those in the Image Recording Guide - Exclusive Lip-sync.

The 4K Version differs from the Studio Avatar in material format in only two aspects:

1. The video size should not exceed 10 GB.
2. The video resolution should be 4K (3840 x 2160), with an aspect ratio of 16:9 (or 9:16).

Voice Clone Recording Guide - Basic Edition

Last updated：2024-09-18 20:42:48

I. Custom Material Self-Check Items

For voice clone, you need to submit an audio recording containing 100 sentences. Before submission, ensure that you check each of the following self-check items:

1. Ensure no other voices are recorded besides the voice of the person being cloned.
2. The audio recording should have a moderate volume, with no noticeable reverb, background noise, or other disturbances.
3. Record using Mandarin Chinese; the text should be diverse, with no excessive repetition of sentences.

Audio format requirements:

1. All audio files must be converted to WAV format and submitted as a compressed ZIP package.
2. Directly select all audio files and compress them into a ZIP package (do not create a new folder before compressing). The ZIP file should not exceed 1 GB.
3. Each audio file must have a sampling rate of 24 kHz or higher, and the length of each file should not exceed 1 hour.
4. Audio file names should not contain spaces or special characters.

II. Audio Recording Guide (Text Version)

Recording content

Follow a pause-read-pause cycle, reading 100 sentences in sequence and generating the audio.

Recording text: You may choose text from your field of expertise, or see the attachment reference texts. The more sentences you include, the better the training results will be.

Text requirements: The text must be in Chinese characters. Individual sentences should not exceed 50 characters, with an average sentence length of around 20 characters.

Number of audio files: The recording can be a single continuous segment or divided into multiple segments, with a maximum of 10 files.

Audio format: It is recommended to use lossless WAV format for recording (specific formats are not restricted), with a sampling rate of no less than 24 kHz.

Notes

The environment should be quiet with no background noise. It is recommended to use a microphone with a windscreen, keeping it within 10 cm of the mouth, and maintaining a moderate volume.

Avoid recording in rooms with smooth walls or floors, such as large glass walls or marble floors, to prevent introducing reverb.

Familiarize yourself with the text before recording to avoid interruptions or disjointed reading.

Be careful to avoid microphone popping.

Pause naturally at the end of each sentence; during the sentence, pause naturally according to the text's normal flow.

Read with rhythm and intonation that reflects your natural speaking style.

Articulate clearly, ensuring that all pronunciations are accurate.

Avoid making any other movements while speaking to prevent unnecessary noises (e.g., clothing rustling and swallowing voice).

Note:

The quality of the custom audio is closely tied to the original recording. High-quality audio will result in a better voice clone, while poor audio quality will lead to a subpar final result.

For example, if the original audio contains noise, the final customized output will also include that noise.

III. Typical Issues

Popping voice

Avoid popping voice, which typically occurs when the microphone is too close, lacks a pop filter, or the recording volume is too high.

Lip smacks, saliva noises, breathing, and microphone pops

Avoid excessive lip smacking, saliva noises, and noticeable breathing voice caused by frequent mouth opening and closing or swallowing during the recording process. Minimize microphone pops as well.

Noise and reverb

Avoid placing the microphone too far from the mouth and recording in environments with significant background noise, such as voices, air conditioning, or background music. Also, avoid introducing reverb, which is often pronounced in rooms with many glass surfaces or smooth walls.

Missing frequency spectrum

Avoid using recording software with built-in enhancement or noise reduction modules, as these can damage the original audio and result in missing frequency bands in the spectrum.

IV. Audio Quality Detection Interface Specification Explanation

Currently, the Audio Quality Testing Task Creation API allows you to detect the following metrics, which help identify issues within the audio. The metric descriptions are as follows:

Signal-to-noise ratio (SNR): The ratio of the useful signal energy to the noise energy in the audio. The higher, the better. An SNR of 30 or above is considered acceptable.

Causes of low SNR:

This may be due to a noisy recording environment. Consider recording in a quieter location.

It could also be due to the mouth being too far from the microphone, resulting in insufficient useful signal energy. Adjust the distance between the microphone and the mouth to around 10 cm. (Being too close may cause microphone pops or clipping.)

Reverberation index: The ratio of useful signal energy to echo energy in the audio. The higher, the better. A value of 30 or above is considered acceptable.

Causes of low reverberation index:

It may be due to an unsuitable recording environment that produces echoes. Large spaces or hard walls can easily generate echoes. Try to record in smaller spaces with more soft surfaces, such as a bedroom or inside a car.

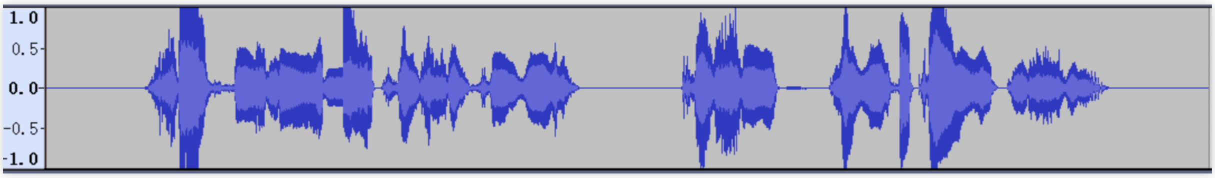
Clipping: Clipping indicates that parts of the audio exceed the maximum allowable amplitude, which, simply put, means the audio volume is too high. A value of 0 or less is considered acceptable.

Causes of clipping:

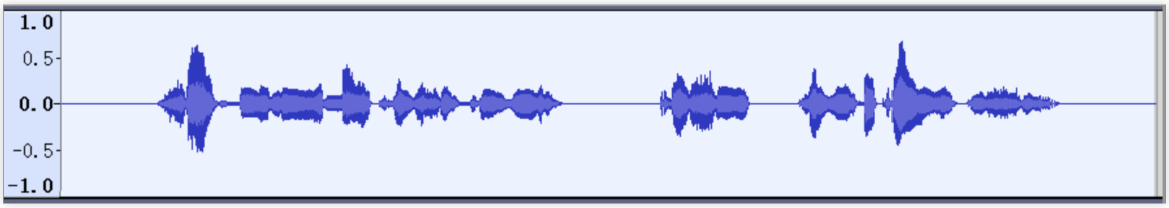
This is typically caused by the mouth being too close to the microphone during recording. Adjust the distance between the microphone and the mouth to about 10 cm.

It may also be due to the recording software's volume setting being too high. This can be resolved by lowering the volume in the recording software.

Waveform illustration of clipped audio:



Waveform illustration of audio with no clipping:



Partial audio examples:

The attachment includes example audio clips labeled as High-Quality Audio, Reverberation Not Meeting Standards, Signal-to-Noise Ratio Not Meeting Standards, Both SNR and Reverberation Not Meeting Standards, and Audio with

Clipping. These are available for download and listening.

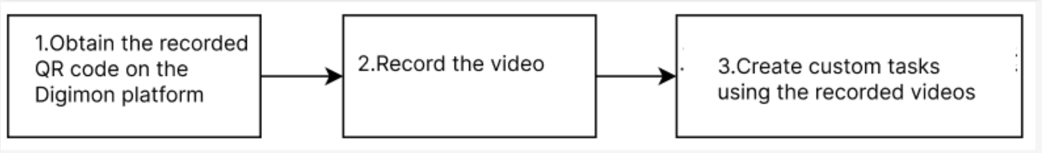
Audio_examples.zip(1.1MB)

下载

Voice Clone Recording Tool - Basic Edition

Last updated：2024-09-18 20:43:05

We support WeChat QR code scanning for recording training data on mobile devices, making it easy to create a voice clone task. The customization process is divided into the following three steps:

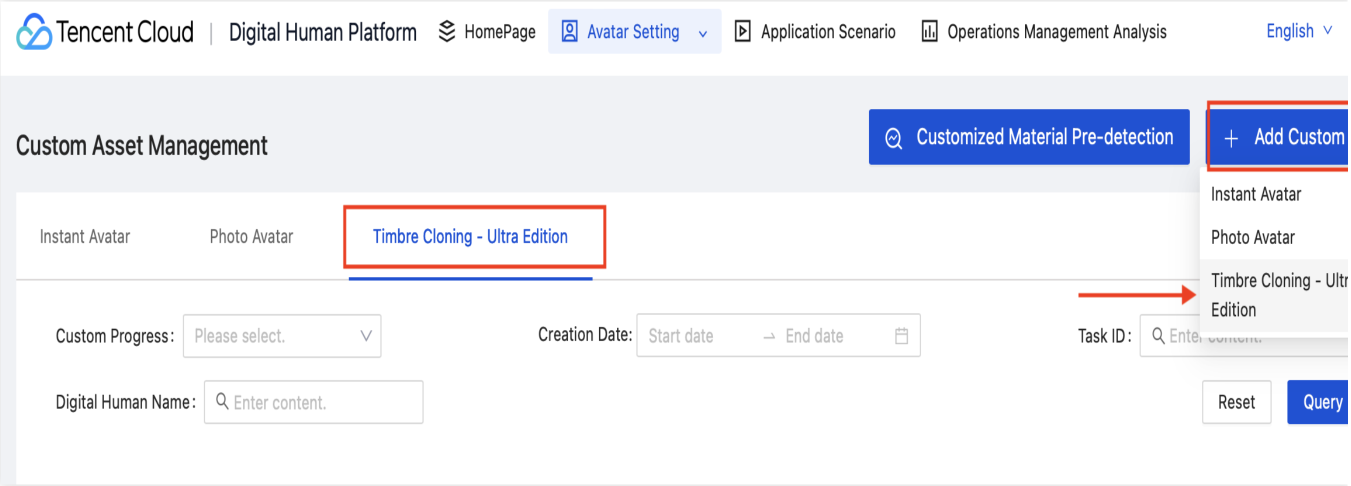

1.Obtain the recorded QR code on the Digimon platform

2.Record the video

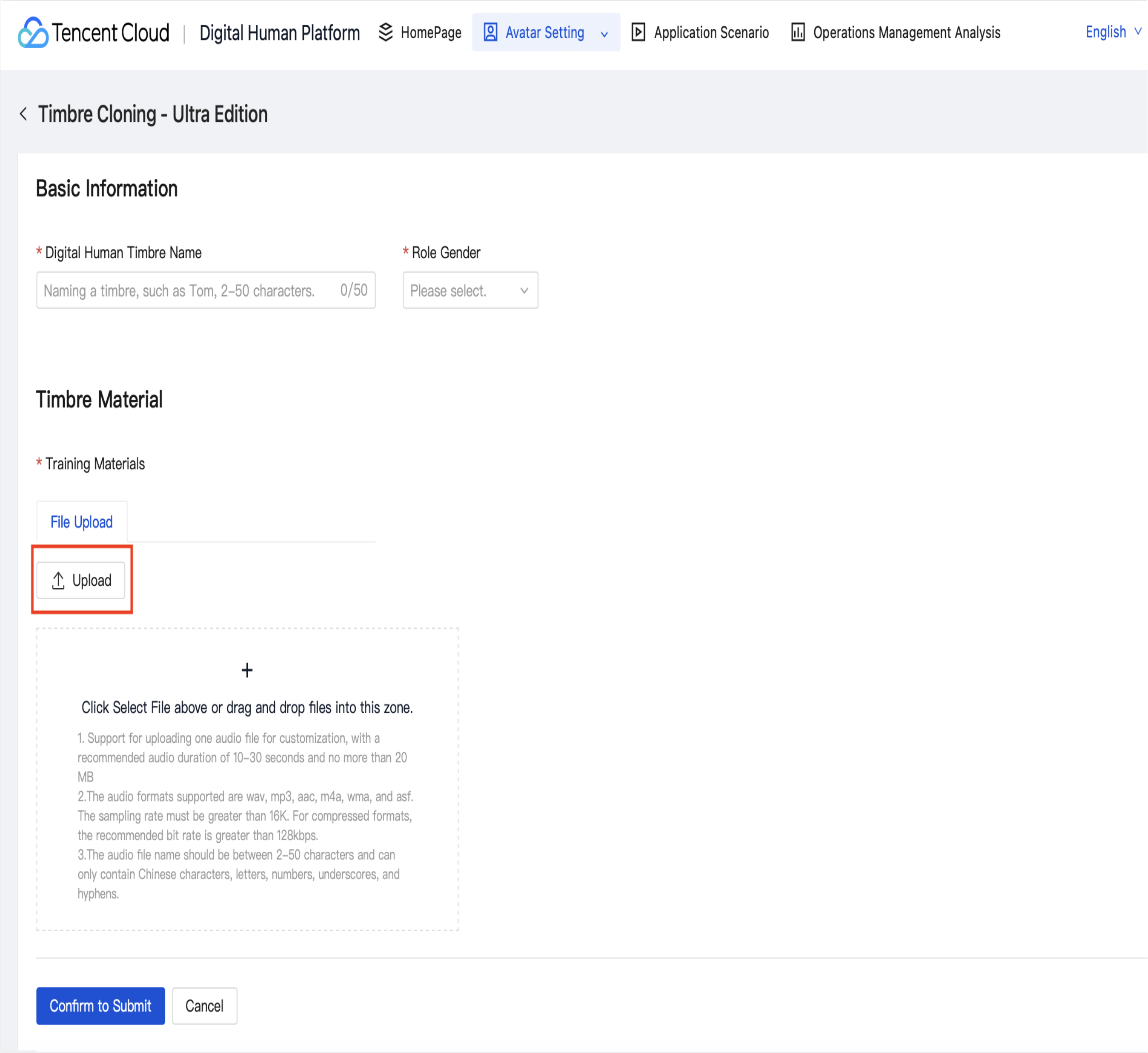
3.Create custom tasks using the recorded videos

Use this material on the Digital Human Platform to create a custom task

1. Log in to the Digital Human platform and create a voice clone customization task. The entry point is located at Digital Human Platform → Image Settings → Custom Asset Management → Add 2D Small Sample Customization, as shown in the figure below:


Tencent Cloud | Digital Human Platform
HomePage
Avatar Setting
Application Scenario
Operations Management Analysis
English
Custom Asset Management
Customized Material Pre-detection
+ Add Custom
Instant Avatar
Photo Avatar
Timbre Cloning - Ultr Edition
Instant Avatar
Photo Avatar
Timbre Cloning - Ultra Edition
Custom Progress: Please select.
Creation Date: Start date → End date
Task ID: Enter content.
Digital Human Name: Enter content.
Reset
Query

2. When the training material is uploaded, select participants submitted materials through the Recording Tool. As shown in the figure below:


Tencent Cloud | Digital Human Platform
HomePage
Avatar Setting
Application Scenario
Operations Management Analysis
English

Timbre Cloning - Ultra Edition

Basic Information

* Digital Human Timbre Name: Naming a timbre, such as Tom, 2–50 characters. 0/50

* Role Gender: Please select.

Timbre Material

* Training Materials

Click Select File above or drag and drop files into this zone.

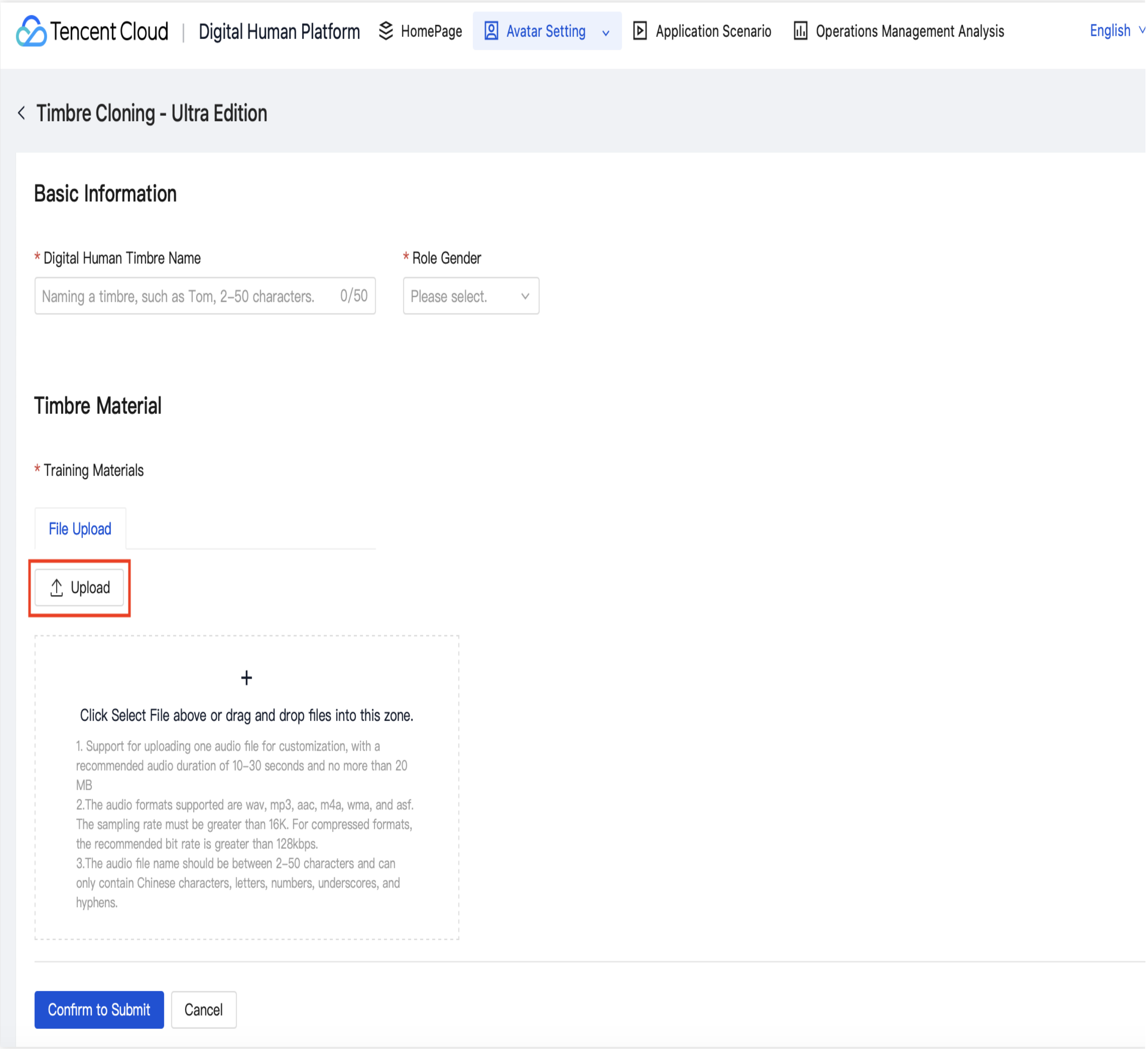
1. Support for uploading one audio file for customization, with a recommended audio duration of 10–30 seconds and no more than 20 MB
2.The audio formats supported are wav, mp3, aac, m4a, wma, and asf. The sampling rate must be greater than 16K. For compressed formats, the recommended bit rate is greater than 128kbps.
3.The audio file name should be between 2–50 characters and can only contain Chinese characters, letters, numbers, underscores, and hyphens.

Voice Clone Recording Guide - Ultra-fast Version

Last updated：2024-09-18 20:43:27

After purchasing quotas, you can use the platform to directly record material for voice clone.

Access path: Homepage > **Image Setting** > **Customized Asset Management** > **Add New Custom Task** > **Timbre Cloning - Ultra Edition**, as shown in the figure below.



Timbre Cloning - Ultra Edition

Basic Information

* Digital Human Timbre Name: Naming a timbre, such as Tom, 2–50 characters. 0/50

* Role Gender: Please select.

Timbre Material

* Training Materials

Click Select File above or drag and drop files into this zone.

1. Support for uploading one audio file for customization, with a recommended audio duration of 10–30 seconds and no more than 20 MB
2.The audio formats supported are wav, mp3, aac, m4a, wma, and asf. The sampling rate must be greater than 16K. For compressed formats, the recommended bit rate is greater than 128kbps.
3.The audio file name should be between 2–50 characters and can only contain Chinese characters, letters, numbers, underscores, and hyphens.

You can also submit the material for customization through APIs: see Interface Call Logic Diagram for details.

Voice Clone Recording Guide - Ultra-Fast Version (Minority Language)

Last updated：2025-04-14 14:42:07

Before integration, you can check our supported language list:Appendix 4 - Language List.

Preparations (Purchase Quota and Training Material)

After purchasing quotas, you can use the Digital Human Platform to directly record material for multilingual voice clone.
Access path: homepage > image settings > custom asset management > add custom task > voice clone (ultra-fast version - minority language), as shown below.

You can also submit the material for customization through APIs: see Interface Call Logic Diagram for details.
The main information to fill in includes: defining the timbre name, determining the gender of the timbre, and selecting the language for training.
The mainly uploaded materials include: authorized audio (upload after recording according to the specified content. Note that you need to strictly abide by the requirements here. There will be related prompts on the page) and audio materials that need to be trained.

The audio requirements are as follows:
1. Supports uploading 1 audio file for customization. The recommended audio duration is 10 - 90 s, no more than 20 M;
2. Audio format support: wav, mp3, aac, m4a, wma, asf; Sampling rate support: 16K, 24K, 48K; For compression format, bitrate higher than 128 kbps is recommended;
3. The audio name should be 2 - 50 characters long. Only Chinese characters, letters, digits, underscores and hyphens are allowed.

Submit Materials, Enter Training

After all materials have been transmitted, click "Confirm Submission". The following pop-up will appear. Select "Agree and Submit". Under normal circumstances, the voice type will enter the training status.

View Training Process

After submission, a notification will pop up: Submission succeeded (as shown above). On this page, you can directly click "view progress" to navigate to the Progress Query page. You can also directly click to view the position shown below to check the training progress of the voice type. When the display is completed, you can use this voice type in " Application Scenario".

Note:

If the Customized Text To Speech fails, don't worry. The related quota will be automatically returned and you can continue to retry the training.

Custom Material Submission Guide

Last updated：2025-04-08 17:53:48

Submitting through the API Interface

See API document 2D Real Person Small Sample Avatar, Voice Customization API Document.

Submit Via the Platform

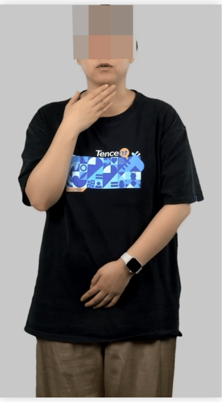
See Custom Asset Management in the platform operation guide.

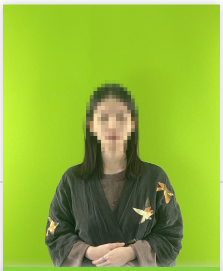
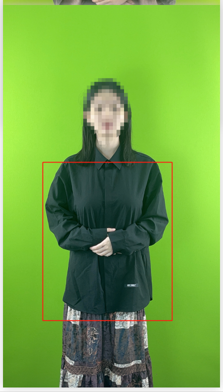
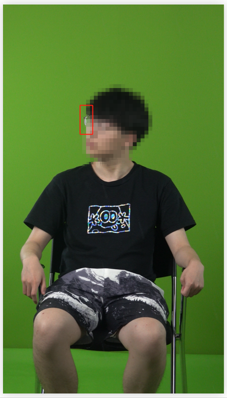
FAQs

Last updated：2024-09-18 20:44:07

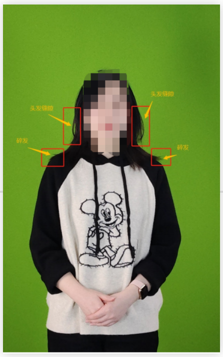
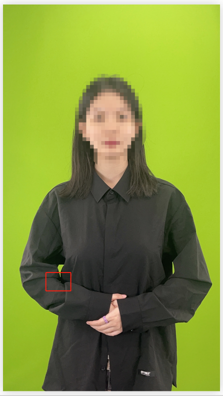
I. Issues Related to Avatar Customization

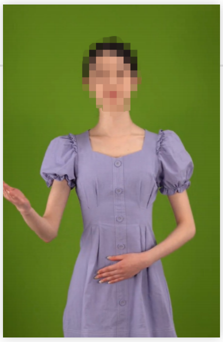
Issue Description	Answer	Optimization Suggestions	Example
What impact will there be if there are edits or frame skips in the middle of the video material?	The generated Avatar will experience frame skips at the same position. During the demo export, you can manually select segments, but you need to ensure there is at least 1 minute of continuous footage.	The video can only be edited by trimming the beginning and the end. The provided material must be continuous and uninterrupted footage.	
What impact will there be if beauty effects in the video material cause shaking (face, waist, etc.)?	The generated Avatar will experience shaking at the same position. During the demo export, you can manually select segments, but you need to ensure there is at least 1 minute of continuous footage.	If the video requires beautification, it is recommended to check the processed footage for any shaking.	
What impact will there be if the video material shows the person turning their head at a large angle, with a pronounced side profile, or looking up and down?	1. This can cause face detection to fail during training, resulting in an effective video duration of less than 3 minutes, which may negatively impact the final lip-syncing effect. 2. The lip-syncing effect will be significantly worse before and after large head turns.	It is recommended that the person’s head does not make large turns. When filming at an angled sitting position, ensure that the mouth is fully visible.	
What impact will there be if the face	1. Obstructions to the face can cause the training to	Ensure that hands do not enter the head	

or chin is obstructed by hands or other objects in the video material?	end prematurely, resulting in an effective video duration of less than 3 minutes, which may negatively impact the final lip-syncing effect. 2. If the chin is obstructed, the resulting Avatar may have missing elements where the chin is covered, such as the hand being overlapped by the mouth during lip-syncing.	area while making movements; the face should remain fully visible without interruptions throughout the video.	
Video duration is too short, less than 3 minutes. Video duration is insufficient, extended by repeating and stitching segments together. What impact will there be if the video repeats reading the same text segment?	Insufficient variety in lip movements can significantly affect the accuracy of the lip-syncing effect.	It is recommended that the video recording meets a duration of 3 to 5 minutes.	
What impact will camera shaking have?	The generated Avatar may also exhibit shaking. You will need to manually select stable, continuous segments for training. If there is no stable segment of at least one minute, re-recording will be necessary.	Ensure that the camera remains fixed and stable throughout the entire recording, with no changes in the camera position.	
What impact will it have if clothes, chairs, tables, and other accessories are the same color	1. It can easily result in poor segmentation effects. 2. During green screen removal, accessories of the same color may be	Avoid wearing green-toned clothing and using green-toned props.	

as the green screen?	affected, leading to color changes.		
How do I handle large reflections of green light from characters, tables, and props?	During training, applying a higher level of green screen removal may cause color distortion in areas where the green is removed. If a high level of green screen removal is not applied, the resulting Avatar may also exhibit green spill effects.	Optimizing the filming process.	
How do I handle the reflections and green tints in my glasses?	Green areas on glasses may be detected and segmented as background.	During filming, make appropriate adjustments to avoid green reflections in the eyeglass lenses.	
How can I reduce the reflection of green light onto a person?	1. Choose Oxford fabric for the green screen and install backlighting to ensure no light seeps through from behind. 2. Keep the model at least 1.5 meters away from the green screen. 3. Surround the model with lights to eliminate green tints on the contours. 4. Light the green screen and model separately. 5. Use black cloths to cover surrounding green areas that are not needed		

	for the shot, reducing diffused light.		
What impact will there be if there are multiple voices speaking in the video?	It can interfere with lip movement recognition, negatively affecting the quality of the lip-syncing.	Ensure a quiet environment during recording. If this is not possible, adjust the microphone's pickup range to minimize the capture of other voices or apply post-processing to the audio.	
What impact will there be if a blue screen is used for recording?	Blue light reflected on the body cannot be removed; you will need to manually key out the blue light and provide the processed material along with the channel file. This will allow for normal production.	Prepare a green screen in advance to avoid using a blue screen for recording.	
What impact will there be if the gaze does not focus on the camera and is unsteady?	The generated Avatar's gaze will also appear unsteady and unfocused.	It is recommended to maintain direct eye contact with the camera throughout the entire recording.	
What impact will there be if the recording does not start with 3 seconds of silence?	1. It may cause the generated Avatar's mouth to remain open during silent moments. 2. During training, you can manually select frames where the mouth is closed between speech segments to use as silent frames, but this may result in unnatural transitions.	It is recommended to start with 1-3 seconds of silence, keeping the mouth in a closed position.	
What impact will there be if hair partings and stray	1. Stray hairs outside the main area of the hair may disappear during segmentation.	Before recording, fix your hair to ensure that no parting is visible against the	

hairs are prominent?	2. A prominent parting may cause the outer hair to disappear or flicker, and the parting itself may not be segmented properly.	green screen and minimize stray hairs as much as possible.	
What impact will there be if earrings are worn?	1. If the background around the earrings is a green screen, the earrings may disappear or flicker after segmentation. 2. If the area around the earrings is covered by hair, there will be no impact, and the segmentation will proceed normally.	It is recommended to avoid wearing earrings. If earrings are worn, ensure that they remain within the area covered by hair in the frame.	
What impact will there be if metal accessories, such as jewelry, buttons, watches, or necklaces, are worn?	1. They may reflect the green screen, and after green screen removal, this could result in color changes or flickering. 2. A necklace that is too thick or positioned near the head area may cause face recognition to fail.	It is recommended to avoid wearing metal accessories. If you choose to wear them, ensure that the accessories minimize green screen reflections.	
What impact will there be if there are small closed gaps under the arms or between the legs?	In cases where the gaps are too dark or extremely small, segmentation may not be accurate.	Adjust the pose during filming to either increase the gap or ensure no gaps are visible.	
What impact will there be if movements extend beyond the frame?	When the generated Avatar performs the same action, the parts that go outside the frame will disappear. During training, you will need to manually select continuous segments where the	Keep movements within the frame, avoiding any actions that extend beyond the boundaries of the screen.	

	actions stay within the frame.		
What impact will there be if there are many specific actions with low reusability (e.g., gestures like showing numbers 1, 2, 3)?	When the generated Avatar performs random actions, these specific gestures may not match the text content, leading to unnatural results.	When performing actions, try to use more general, versatile gestures.	
What impact will there be if the video has significant background noise?	It may negatively impact the accuracy of the lip-syncing effect.	It is recommended to use a microphone for recording. You can also lower the recording volume and increase the speaking volume accordingly.	
What should be considered when recording a seated posture using a table?	Ensure that the table does not reflect green from the green screen and that it remains stable without any shaking.		
Can side-profile recording be done?	The facial features and mouth must remain fully visible at all times. Avoid turning too far to the side.		
Is it necessary to record audio when capturing the Avatar?	Yes, it is absolutely necessary, and audio and video must be synchronized. The algorithm requires audio and video to form a paired set for lip-sync training, so the audio corresponding to the video is essential.		
What should be done if a fill light or other objects appear in the frame?	1. Ensure that the person is fully within the green screen area. 2. Ensure that any unnecessary objects do		

	not overlap or intersect with the body in the frame, maintaining a clear separation.		
What should be done if the text is read incorrectly?	1. Mispronounced words during the video recording process can be ignored. 2. If a word is mispronounced during audio recording, pause for two seconds and then re-read the sentence.		
Can the client's own text be used?	Yes, and it is recommended that the client reads text that aligns with the type of content being produced.		

II. Issues Related to Voice Clone

Issue Description	Answer	Optimization Suggestions
What impact do reverb and noise have?	This can easily lead to poor results after voice training.	1. Choose a room with minimal echo and good voiceproofing (e.g., a bedroom) for recording. 2. Use a microphone for recording, and adjust the microphone settings to minimize noise pickup. 3. Enhance the demo quality by applying post-production audio processing to reduce reverb and noise.
What impact will there be if the Mandarin pronunciation is not standard?	After voice training, the pronunciation may sound unusual.	It is recommended to use standard Mandarin with clear enunciation.
What impact will there be if the ASR segmentation contains fewer than 50 sentences?	If the overall recording duration is short, resulting in fewer sentences, it can severely affect the voice quality, and additional recording will be necessary.	Follow the voice recording guidelines to ensure the recording contains at least 100 segments and exceeds 10 minutes in duration.

What happens if the audio amplitude is too high (popping)?	The trained voice may also exhibit the same pronunciation issues.	You can debug the recording equipment to improve the audio quality or provide post-processed audio material.
What impact will there be if the audio contains noticeable saliva voice or breathing noises?	The trained voice may also exhibit the same pronunciation issues.	Ensure that you avoid these issues during recording, or provide post-processed audio material.
How should I choose a proper location for audio recording?	It is recommended to record in a quiet location with plenty of soft materials, such as in bedrooms or cars.	