

CHUYÊN ĐỀ: ỨNG DỤNG AI VÀO MẠNG

XÃ HỘI

BÀI 2. AI PHÁT HIỆN VÀ XỬ LÝ CÁC VIDEO, HÌNH ẢNH, BÀI ĐĂNG CÓ DẤU HIỆU TIN GIẢ, BẠO LỰC, 18+

GIỚI THIỆU

Trong bối cảnh mạng xã hội phát triển mạnh mẽ, việc xuất hiện thông tin giả (fake news), nội dung bạo lực và 18+ là vấn đề đáng lo ngại. Để giải quyết, các công cụ AI được phát triển nhằm phát hiện, cảnh báo và hỗ trợ kiểm duyệt nội dung. Trong đó, hai công cụ tiêu biểu là **Zhuque AI** (phát hiện tin giả) và **Sightengine** (kiểm duyệt hình ảnh/video có nội dung bạo lực, 18+).

NỘI DUNG HƯỚNG DẪN

a. Zhuque AI – Công cụ phát hiện dữ liệu do AI tạo ra

- Chức năng: Đây là một AI của Trung Quốc, được giới thiệu với khả năng phân tích tin tức và phát hiện tin giả dựa trên dữ liệu lớn và xử lý ngôn ngữ tự nhiên.
- Ứng dụng: Hướng đến hỗ trợ báo chí, nghiên cứu và người dùng phổ thông trong việc nhận diện tin sai lệch.
- Điểm mạnh: Được nhấn mạnh ở khả năng phân tích nhanh, kết hợp dữ liệu ngữ nghĩa và nguồn gốc để đánh giá độ tin cậy.

HƯỚNG DẪN SỬ DỤNG AI - CÔNG CỤ PHÁT HIỆN DỮ LIỆU DO AI TẠO RA

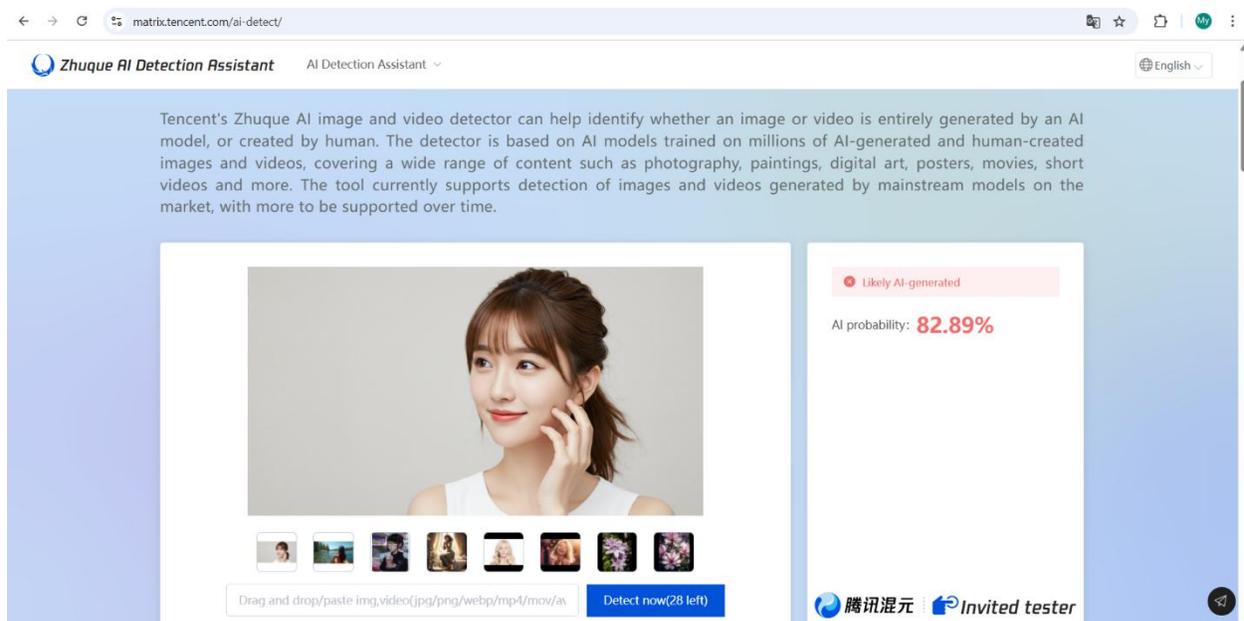
Zhuque AI Detection Assistant

Zhuque AI Detection Assistant

Bước 1:

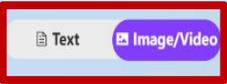
Truy cập trang web: <https://matrix.tencent.com/ai-detect/>.

👉 Đây là công cụ miễn phí của Tencent hỗ trợ kiểm tra văn bản, hình ảnh và video.



Bước 2: Chọn chế độ kiểm tra phù hợp:

- Text → dùng để dán nội dung văn bản cần kiểm tra.
- Image/Video → dùng để tải ảnh hoặc video cần kiểm tra.



Tencent's Zhuque AI image and video detector can help identify whether an image or video is entirely generated by an AI model, or created by human. The detector is based on AI models trained on millions of AI-generated and human-created images and videos, covering a wide range of content such as photography, paintings, digital art, posters, movies, short videos and more. The tool currently supports detection of images and videos generated by mainstream models on the market, with more to be supported over time.



Likely AI-generated

AI probability: **82.89%**

- Nếu chọn Text: copy và dán đoạn văn bản vào ô trống.

Lưu ý: Nội dung văn bản phải từ 250 từ trở lên.

Example 1: AI text

Example 2: AI text

Example 3: human text

Example 4: human text

Khi giới thiệu thuật ngữ tư duy tính toán, Wing [19] mô tả nó như một cách con người để giải quyết vấn đề. Nó bao gồm tập hợp các công cụ tư duy của khoa học máy tính để biến một vấn đề khó thành vấn đề dễ giải hơn. Ủng hộ quan điểm của Wing, Guzdial [16] là một cách suy nghĩ về tính toán. Những người tham gia hội thảo về phạm vi và bản chất của tư duy tính toán [16], mặc dù không được giao nhiệm vụ định nghĩa, vẫn đồng ý rằng nó bao gồm một loạt công cụ và khái niệm từ khoa học máy tính. Ý tưởng này được mở rộng thành việc biểu diễn vấn đề như các quá trình thông tin và giải pháp như các thuật toán [7].



Upload

Clear

Detect now(19 left)

Not likely to be AIGC Report

AI content ratio: **0%**

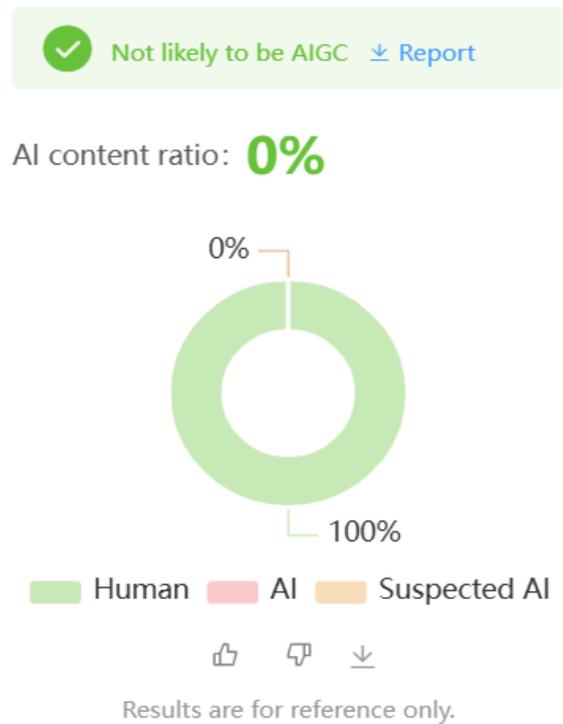


Human AI Suspected AI

Results are for reference only.

+ Xem kết quả hiển thị ở bên phải màn hình:

- AI content ratio 0% (màu xanh lá): Nội dung gần như chắc chắn do con người viết.
- Có phần trăm AI (màu đỏ): Cho thấy nội dung có khả năng do AI tạo ra.
- Suspected AI (màu vàng): Hệ thống nghi ngờ có yếu tố AI.



腾讯混元 : Invited tester

- Nếu chọn Image/Video: tải file ảnh hoặc video lên từ máy tính/điện thoại.

Upload image/video

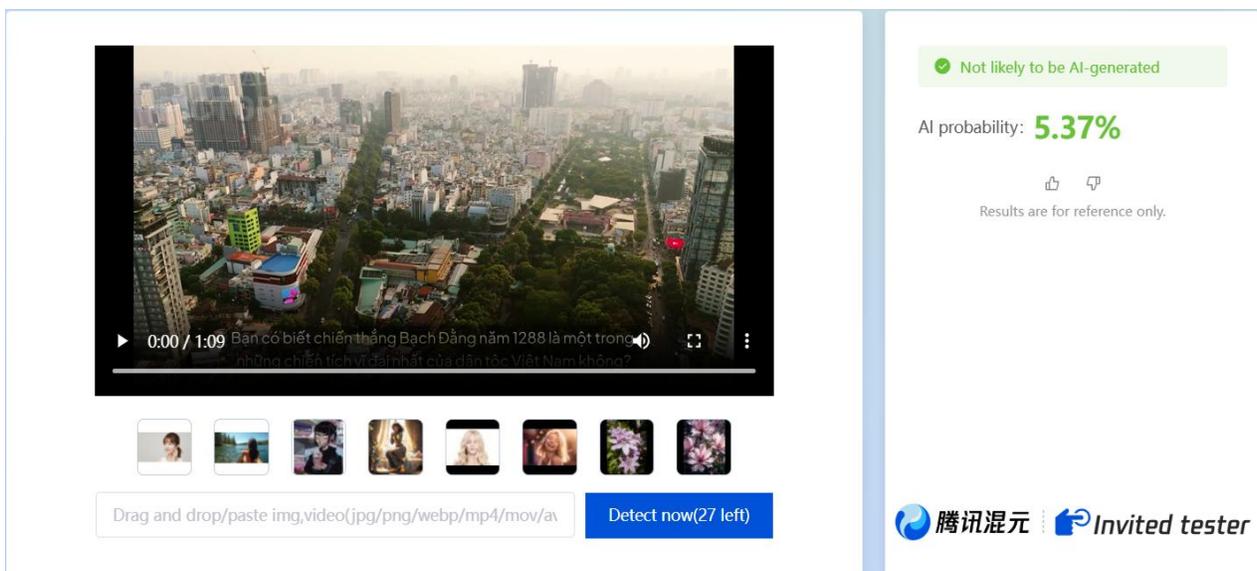
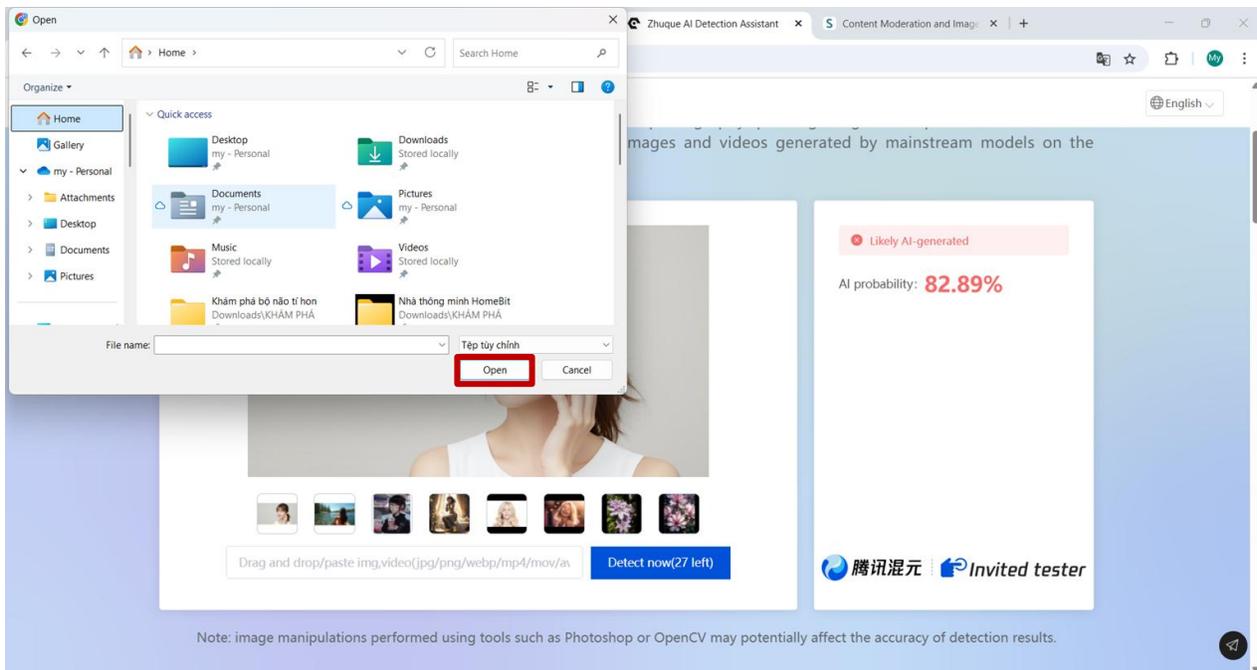
Likely AI-generated

AI probability: **82.89%**

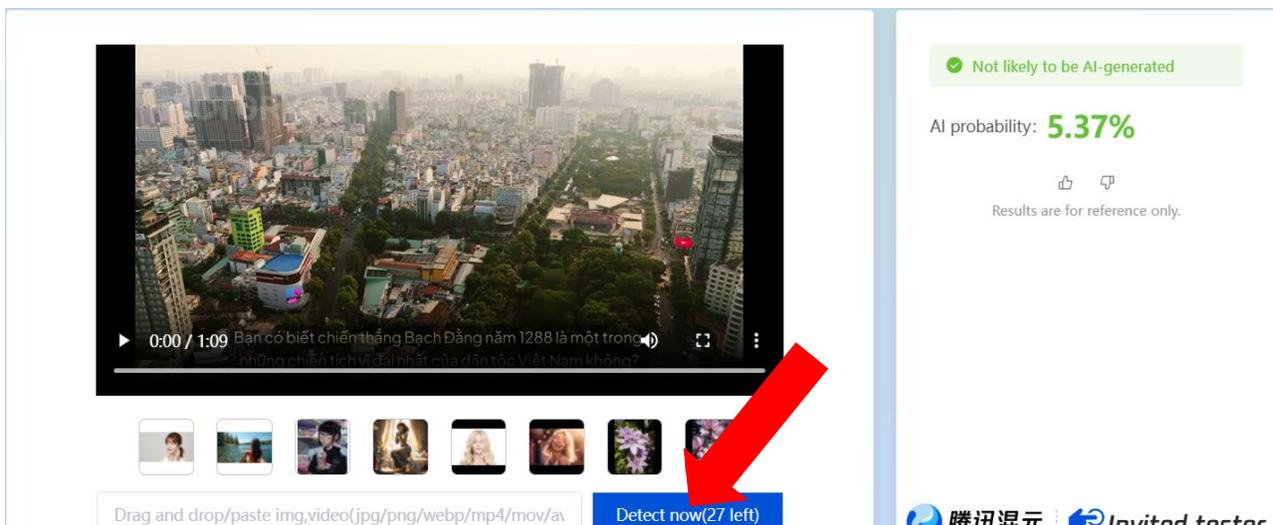
Drag and drop/paste img,video(jpg/png/webp/mp4/mov/avi) Detect now(27 left)

腾讯混元 : Invited tester

The screenshot shows the user interface for uploading an image. A large red arrow points to the 'Upload image/video' text. Below the main image area is a row of eight small thumbnail images. At the bottom, there is a text input field with the placeholder 'Drag and drop/paste img,video(jpg/png/webp/mp4/mov/avi)' and a blue button labeled 'Detect now(27 left)'. On the right side, a pink status bar indicates 'Likely AI-generated' with a red star icon, and below it, the 'AI probability: 82.89%' is displayed in red. The Tencent Huan Yuan logo and 'Invited tester' text are at the bottom right.



Bước 3: Nhấn nút **Detect** để hệ thống xử lý dữ liệu.



Bước 4: Xem kết quả trả về:

- Hệ thống hiển thị xác suất (%) nội dung đó có khả năng được tạo bởi AI.

- Ví dụ:

+ “*AI probability: 5.37% – Not likely to be AI-generated*”

☞ Nghĩa là nội dung này gần như do con người tạo ra.

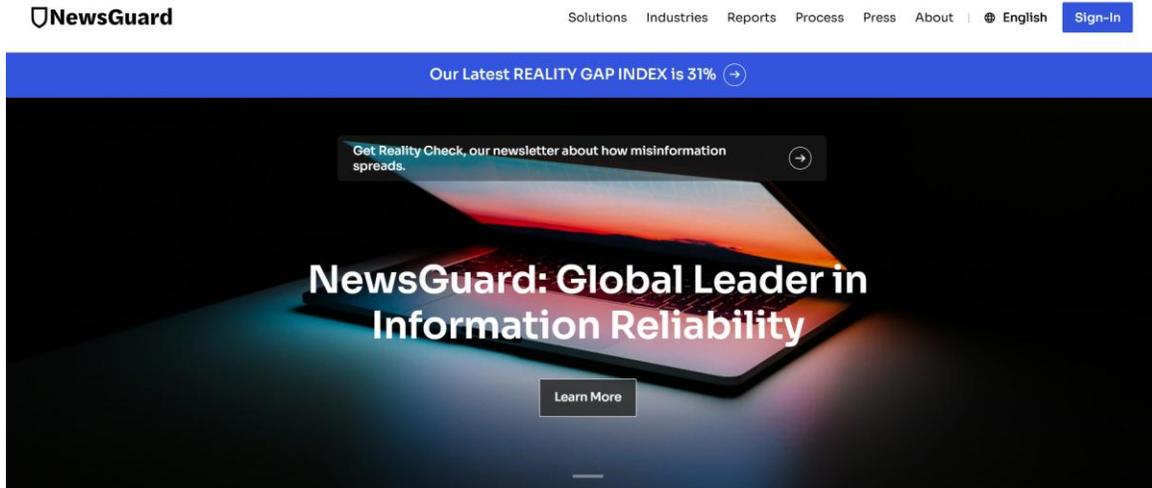


+ “*AI probability: 82.89% – Likely AI-generated*”

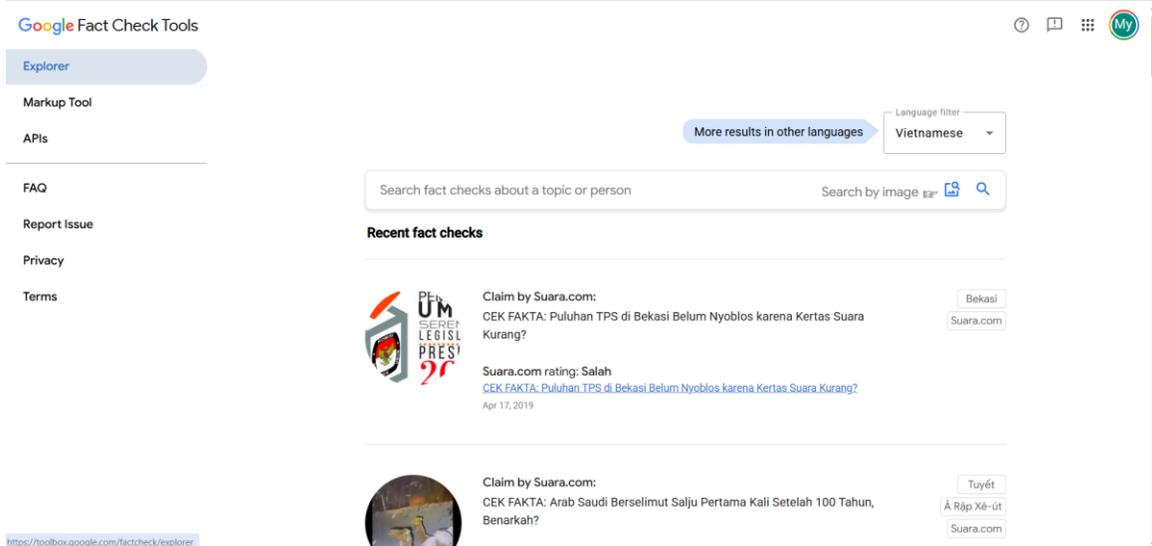
☞ Nghĩa là xác suất nội dung do AI tạo ra rất cao (trên 80%) → hệ thống cảnh báo đây nhiều khả năng là sản phẩm của AI, cần thận trọng khi sử dụng hoặc chia sẻ.

- Gợi ý công cụ AI khác

+ **NewsGuard**: đánh giá độ tin cậy trang tin.



+ **Google Fact Check**: tra cứu tin đã kiểm chứng.

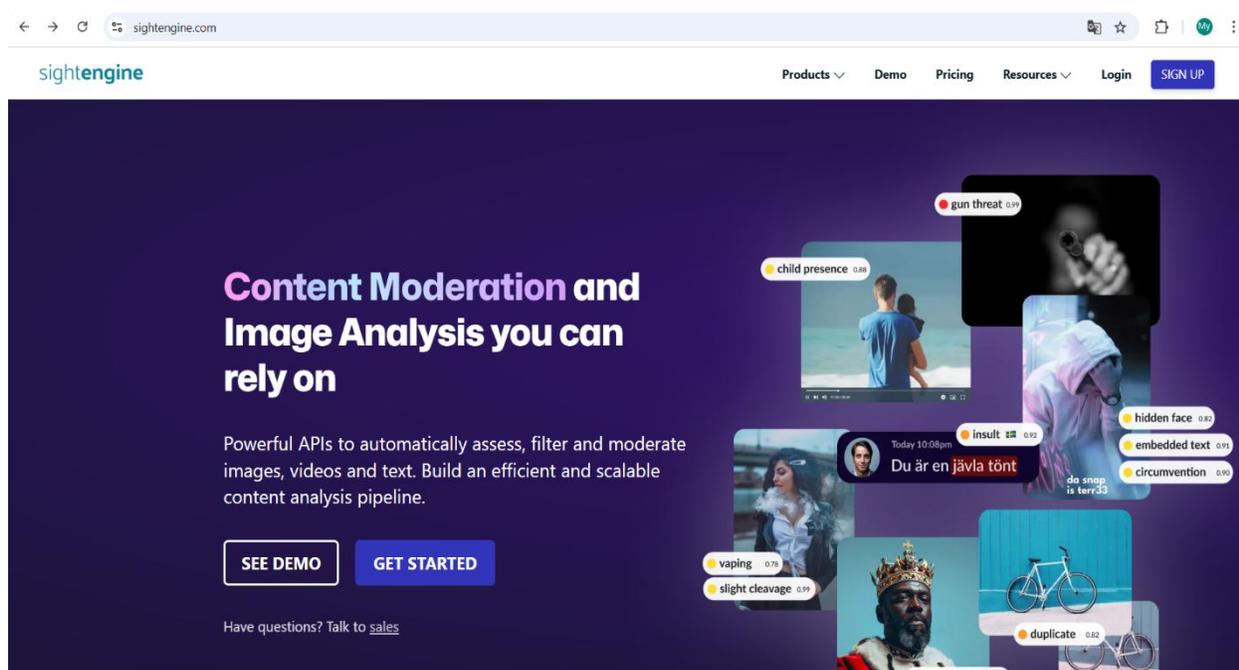


b. Sightengine – Công cụ kiểm duyệt nội dung bạo lực, 18+

- Chức năng: Đây là dịch vụ kiểm duyệt nội dung trực tuyến có thật. Nó phân tích hình ảnh, video và văn bản để phát hiện nội dung nhạy cảm như bạo lực, khỏa thân, khiêu dâm, ma túy, thù ghét...
- Ứng dụng: Thường dùng cho mạng xã hội, app trò chuyện, website hoặc hệ thống quản lý nội dung để tự động gắn nhãn, lọc, hoặc chặn nội dung vi phạm.
- Điểm mạnh: Có API dễ tích hợp, hỗ trợ đa nền tảng (ảnh, video, text), và có hơn 110 lớp phân loại để tăng độ chính xác.

HƯỚNG DẪN SỬ DỤNG SIGHTENGINE – CÔNG CỤ KIỂM DUYỆT NỘI DUNG BẠO LỰC, 18+

Bước 1: Truy cập trang web <https://sightengine.com>.

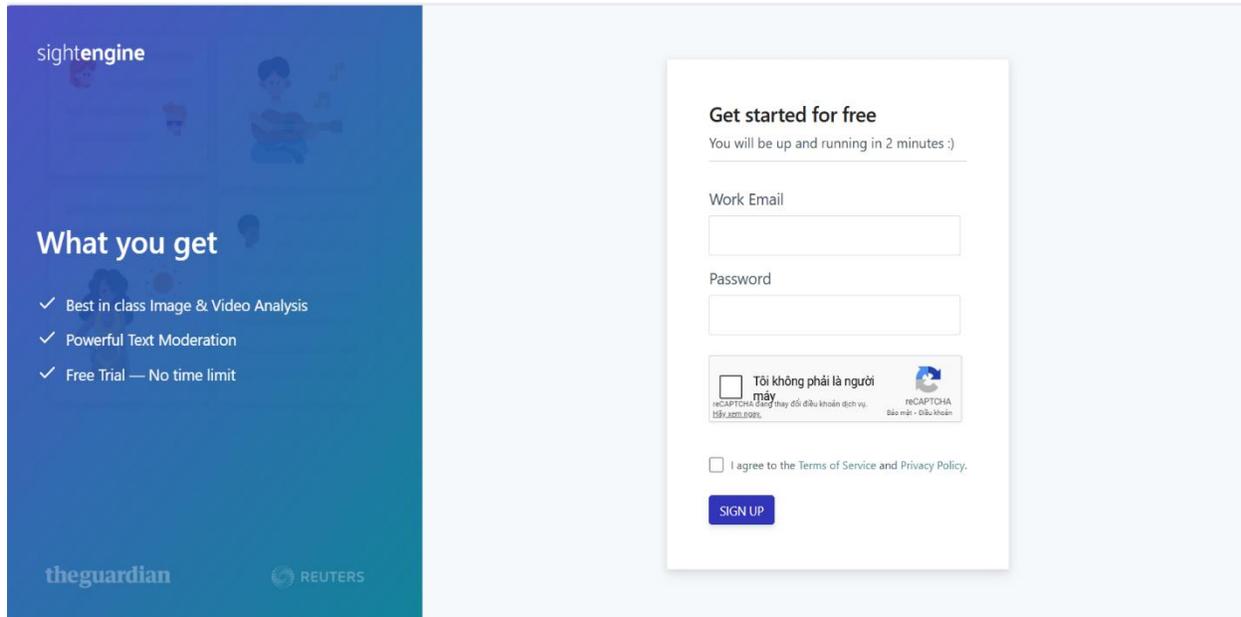


Bước 2: Đăng ký tài khoản miễn phí:

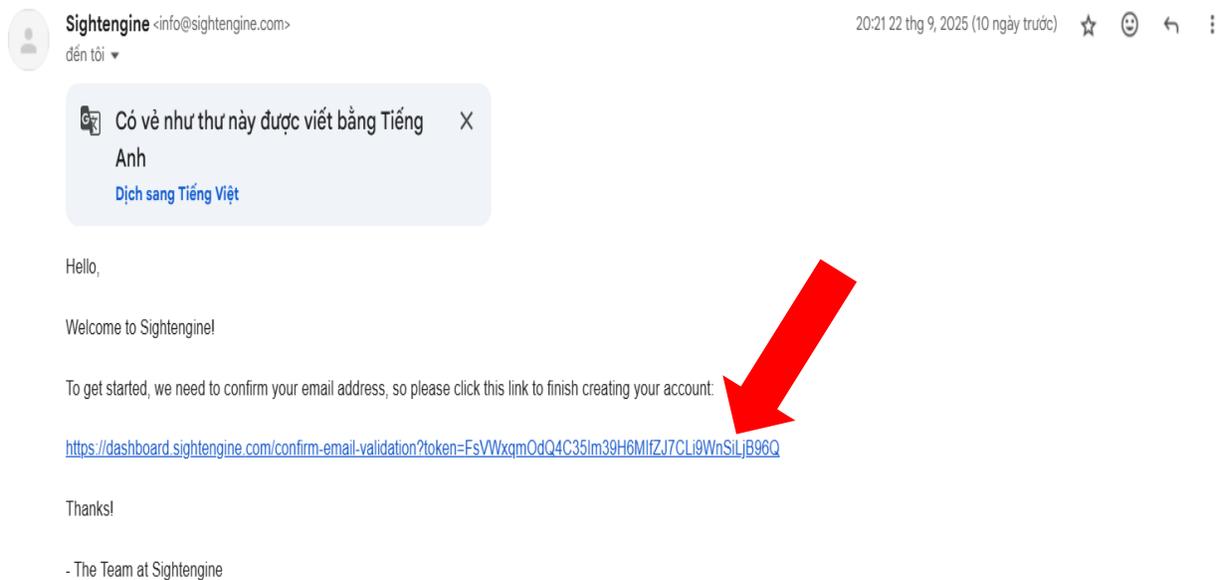
- Nhấn **SIGN UP** ở góc phải.



- Điền email, mật khẩu hoặc đăng ký nhanh bằng tài khoản Google/GitHub.
- Sau khi nhập email, mật khẩu và xác minh, nhấn **SIGN UP**



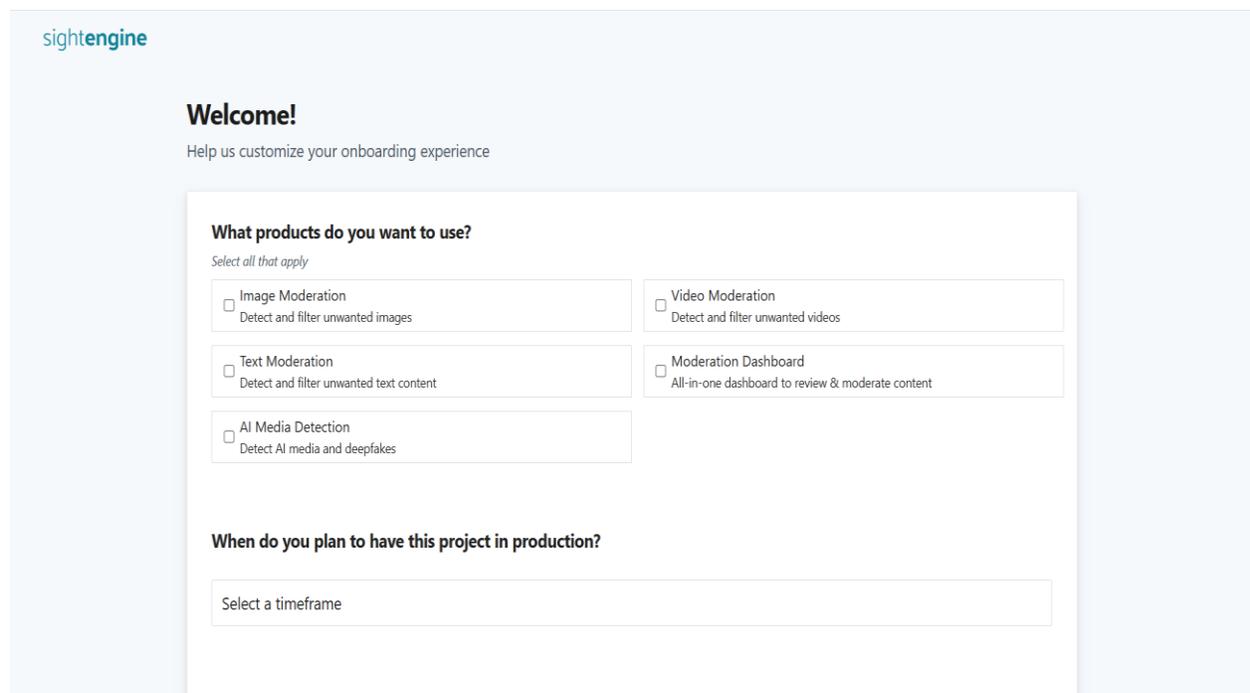
- Tiếp theo, xác minh email bằng cách bấm vào link trong hộp thư.



- Sau khi xác minh email, Sightengine sẽ mở trang **onboarding** để bạn khảo sát dịch vụ cần dùng như hình dưới:

- + **Image Moderation** (lọc ảnh nhạy cảm).
- + **Video Moderation** (lọc video).
- + **Text Moderation** (lọc văn bản).
- + **AI Media Detection** (phát hiện nội dung AI, deepfake).
- + **Moderation Dashboard** (bảng điều khiển tổng hợp).

Bên dưới có mục hỏi **“When do you plan to have this project in production?”** – chỉ là khảo sát thời gian triển khai, có thể chọn tùy ý hoặc bỏ qua.



The screenshot shows the Sightengine onboarding interface. At the top left is the Sightengine logo. Below it is a 'Welcome!' heading followed by the text 'Help us customize your onboarding experience'. The main content area is a white box with a light blue border. It contains a section titled 'What products do you want to use?' with the instruction 'Select all that apply'. There are five checkboxes with labels and descriptions: 'Image Moderation' (Detect and filter unwanted images), 'Video Moderation' (Detect and filter unwanted videos), 'Text Moderation' (Detect and filter unwanted text content), 'AI Media Detection' (Detect AI media and deepfakes), and 'Moderation Dashboard' (All-in-one dashboard to review & moderate content). Below this section is another question: 'When do you plan to have this project in production?' with a text input field containing the placeholder 'Select a timeframe'.

- Cuối cùng nhấn **FINALIZE**.

What products do you want to use?

Select all that apply

Image Moderation
Detect and filter unwanted images

Video Moderation
Detect and filter unwanted videos

Text Moderation
Detect and filter unwanted text content

Moderation Dashboard
All-in-one dashboard to review & moderate content

AI Media Detection
Detect AI media and deepfakes

When do you plan to have this project in production?

As soon as possible

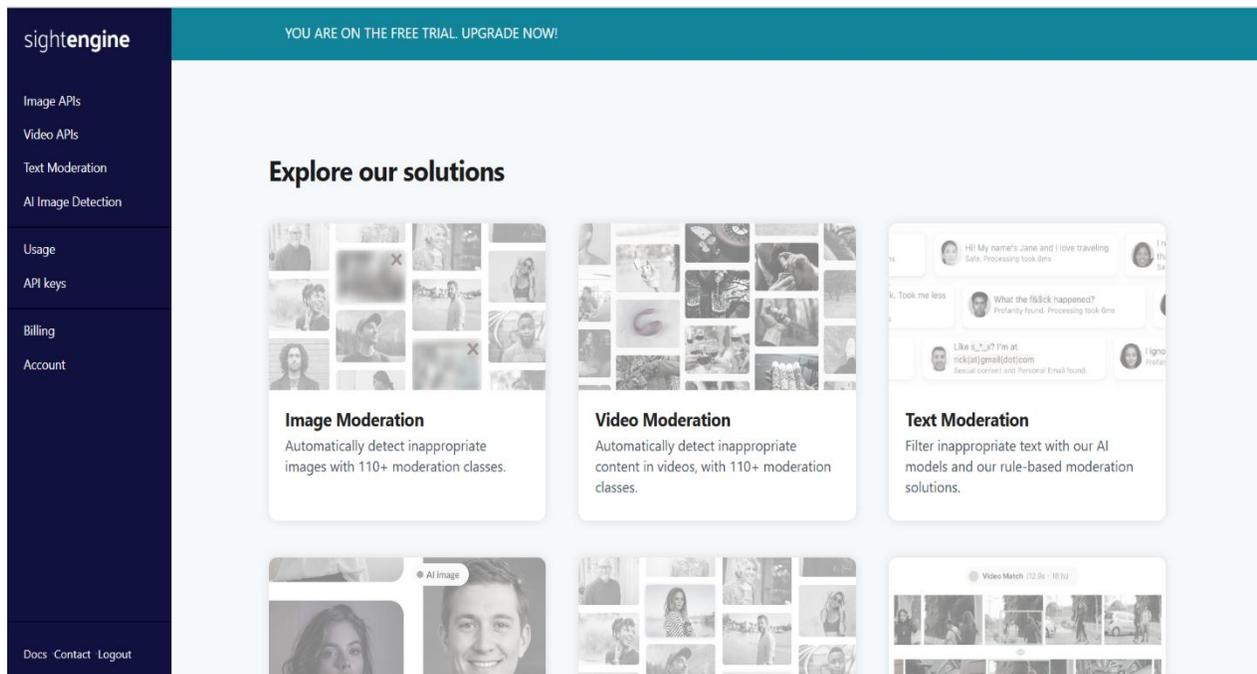
What describes you best?

Trust & Safety

FINALIZE

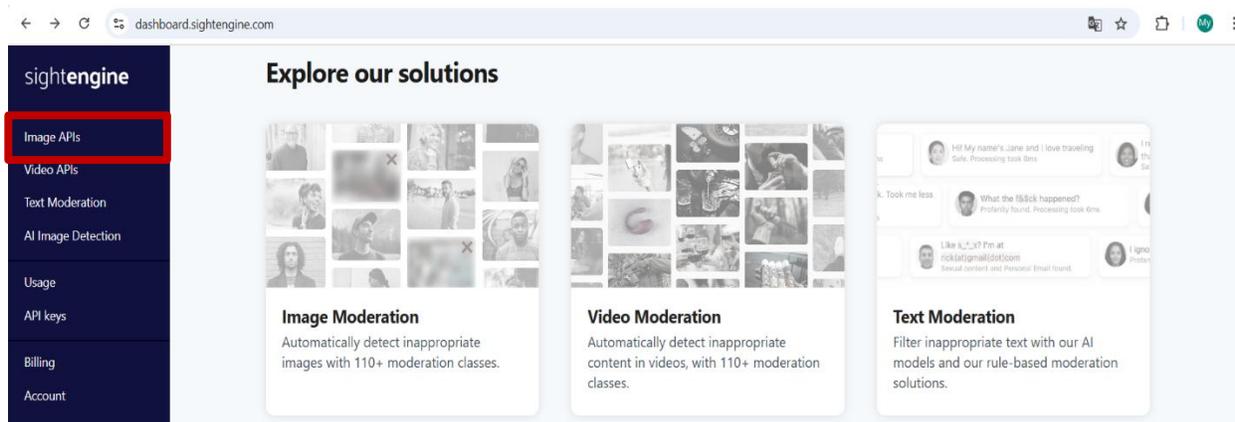
Sau khi đăng nhập, bạn sẽ thấy **Dashboard** với mục **API Keys**.

👉 Đây là phần quan trọng để bạn gửi ảnh/video/văn bản lên hệ thống để phân tích.

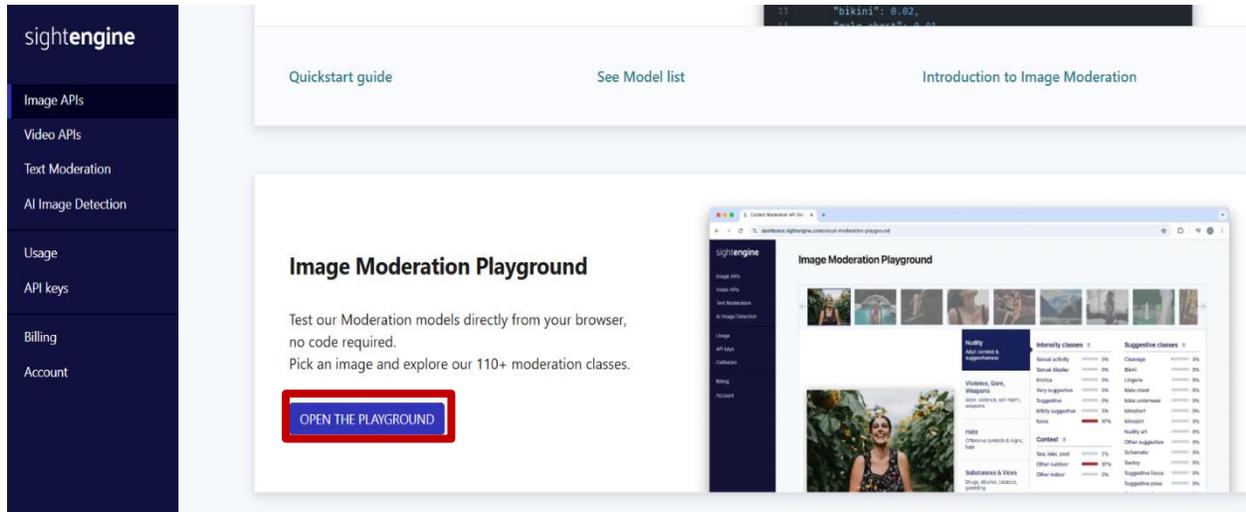


Bước 3: Trong thanh menu bên trái:

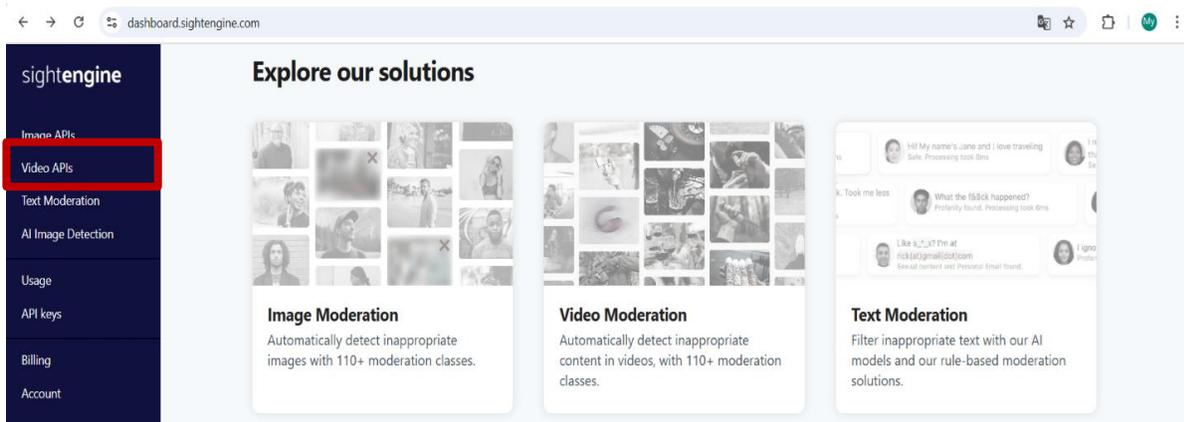
- Chọn **Image APIs** (nếu muốn kiểm tra ảnh).



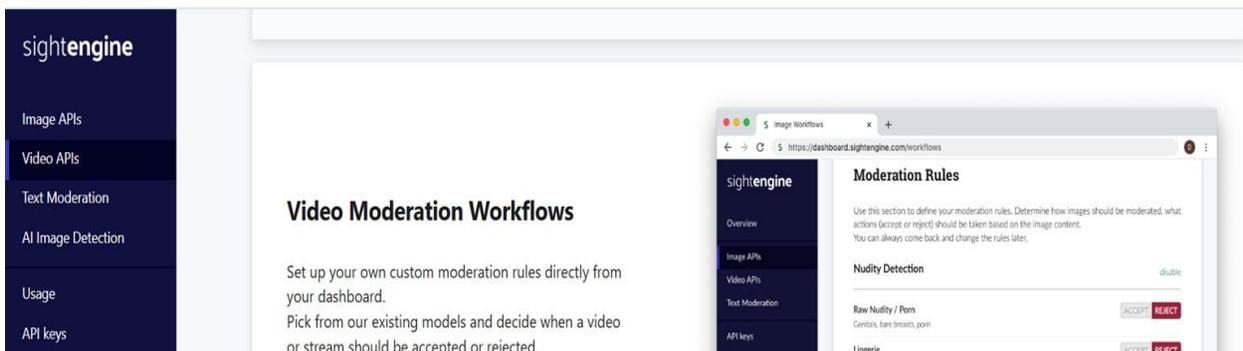
+ Trong Dashboard, tìm mục **Image Moderation Playground** rồi nhấn **OPEN THE PLAYGROUND**.



- Chọn **Video APIs** (nếu muốn kiểm tra video).



+ Trong Dashboard, tìm mục **Video Moderation Playground** rồi nhấn **GO TO WORKFLOWS**.



Bước 4: Upload dữ liệu cần kiểm tra:

- **Ảnh:** Sau khi đã hoàn thành bước 3 hệ thống sẽ hiển thị giao diện như hình dưới.

The screenshot displays the SIGHTENGINE Image Moderation Playground interface. On the left is a dark blue sidebar with navigation options: Image APIs, Video APIs, Text Moderation, AI Image Detection, Usage, API keys, Billing, and Account. The main area is titled "Image Moderation Playground" and features a horizontal strip of image thumbnails at the top. Below this, a large image of a woman in a floral dress is shown. To the right of the image is a detailed analysis panel with the following sections:

- Nudity** (Adult content & suggestiveness): Sexual activity (0%), Sexual display (0%), Erotica (0%), Very suggestive (0%), Suggestive (0%), Mildly suggestive (0%), None (99%).
- Violence, Gore, Weapons** (Gore, violence, self-harm, weapons): 0%.
- Hate & Sensitive Topics** (Offensive, hateful or sensitive content): 0%.
- Substances & Vices** (Drugs, alcohol, tobacco, gambling): 0%.
- Text & QR content**: 0%.
- Intensity classes**: Sea, lake, pool (60%), Other outdoor (40%), Other indoor (0%).
- Suggestive classes**: Bikini (0%), Cleavage (0%), Lingerie (0%), Male chest (0%), Male underwear (0%), Miniskirt (0%), Nudity art (0%), Other suggestive (0%), Schematic (0%), Sextoy (0%), Suggestive focus (0%), Suggestive pose (0%), Swimwear male (0%), Swimwear one (0%).

+ Ở giữa màn hình: hiển thị ảnh bạn vừa tải lên.

+ Bên phải: là bảng phân tích chi tiết với nhiều nhóm chỉ số.

- **Nudity** (nội dung hở thân & gợi dục): liệt kê các mức độ (Sexual activity, Sexual display, Erotica, Suggestive...) → hiển thị % phát hiện.
- **Violence, Gore, Weapons**: dùng để phát hiện bạo lực, máu me, vũ khí.
- **Hate & Sensitive Topics**: nội dung thù ghét, xúc phạm.
- **Substances & Vices**: liên quan đến rượu, thuốc lá, ma túy, cờ bạc.

Ngoài ra, ta có thể dùng các công cụ hỗ trợ dịch thuật được tích hợp sẵn trên trình duyệt web như Google dịch.

daachboard.cinhon.vn/visual-moderation-nlau/vnuvd

Khóa thân
Nội dung dành cho người lớn và gợi ý

Bạo lực, Máu me, Vũ khí
Máu me, bạo lực, tự làm hại bản thân, vũ khí

Chủ đề thù hận và nhạy cảm
Nội dung xúc phạm, thù hận hoặc nhạy cảm

Chất & Tệ nạn
Ma túy, rượu, thuốc lá, cờ bạc

Nội dung văn bản và mã QR
Phân tích văn bản và mã QR

Mô tả nội dung

Các lớp cường độ ?

Hoạt động tình dục	0%
Hiện thị tình dục	0%
Khiêu dâm	0%
Rất gợi ý	0%
Gợi ý	0%
Hơi gợi ý	0%
Không có	99%

Bối cảnh ?

Biển, hồ, hồ bơi	60%
Ngoài trời khác	40%
Trong nhà khác	0%

Các lớp học gợi ý ?

Bikini	0%
Sự phân cắt	0%
Đồ lót	0%
Ngực nam	0%
Đồ lót nam	0%
Minishort	0%
Váy ngắn	0%
Nghệ thuật khỏa thân	0%
Gợi ý khác	0%
Sơ đồ	0%
Đồ chơi tình dục	0%
Tập trung gợi ý	0%
Tư thế gợi cảm	0%
Đồ bơi nam	0%
Đồ bơi một mảnh	0%
Rõ ràng là không mặc quần áo	0%

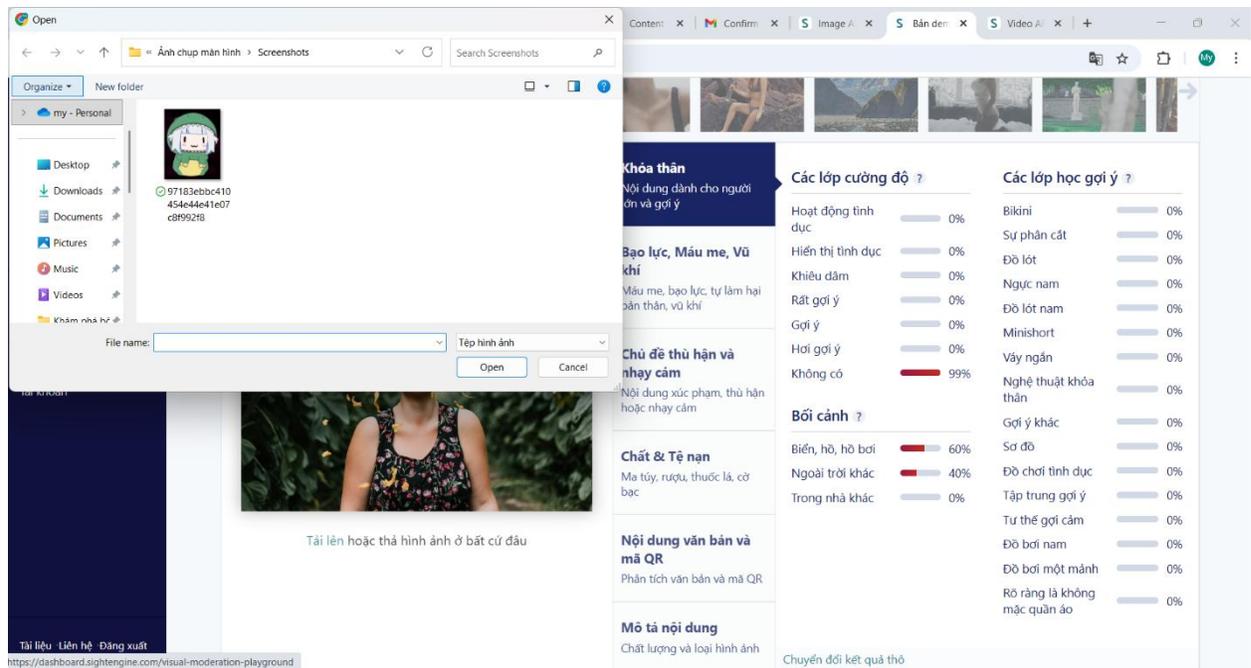
Tải lên hoặc thả hình ảnh ở bất cứ đâu

+ Để tải ảnh lên ta nhấn vào chữ **Upload (Tải lên)** và chọn ảnh muốn tải lên từ máy.



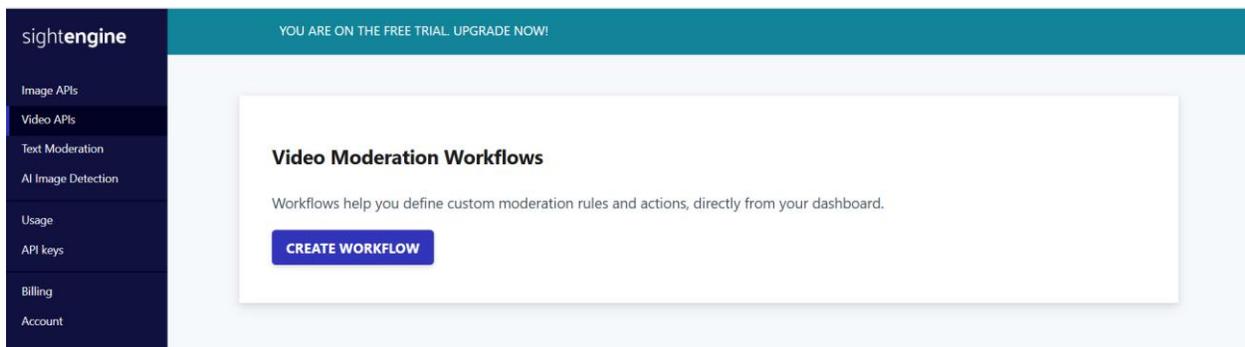
Tải lên hoặc thả hình ảnh ở bất cứ đâu





- **Video:** Sau khi đã hoàn thành bước 3 hệ thống sẽ hiển thị giao diện như hình dưới.

Lưu ý ở đây ta cũng có thể sử dụng Google dịch để hỗ trợ quá trình dịch thuật.



+ Nhấn **CREATE WORKFLOW**.

Cấu hình Workflow (Video Moderation)

- **Workflow name:** Đặt tên rõ ràng (ví dụ *Video_18+_Check*).
- **Moderation mode:**
 - **Fully Automated** → hệ thống tự động quyết định (khuyến dùng để test nhanh).
 - **Hybrid** → cần người duyệt lại một số quyết định.

Workflow information

Workflow name

User videos

Name too short

Automated or hybrid moderation

Determine if all moderation decisions should be automated or if some should be reviewed by human moderators.

Fully Automated Moderation
All decisions are made automatically. No human moderators are involved.

Hybrid Moderation
Some decisions are reviewed/confirmed through your Human Moderation Platform.

+ Ta chọn **Fully Automated Moderation**. Hệ thống sẽ hiển thị phần **Moderation Rules** để bạn chọn loại nội dung cần lọc:

Moderation Rules

Define your moderation rules. Determine what actions (accept, reject...) should be taken based on the video content. You can always come back and change the rules later.

Nudity Detection ENABLE
Detect adult content, nudity and suggestive scenes

Weapon Detection ENABLE
Detect guns, rifles, knives...

Gore & Graphic Violence Detection ENABLE
Detect horrific imagery such as blood, guts, self-harm, wounds, human skulls

Violence ENABLE
Detect violent scenes

Self-Harm ENABLE
Detect intentional self-inflicted injuries or indicators of self-harm

Hate & Offensive Content Detection ENABLE
Detect nazi, supremacist, terrorist, offensive or otherwise hateful imagery

1. **Nudity Detection** → phát hiện nội dung 18+, khỏa thân, cảnh gợi dục.
2. **Weapon Detection** → phát hiện vũ khí (dao, súng, súng trường...).
3. **Gore & Graphic Violence Detection** → phát hiện máu me, vết thương, sọ người...

4. **Violence** → phát hiện cảnh bạo lực (đánh nhau, tấn công).
5. **Self-Harm** → phát hiện hành vi tự gây thương tích, tự hại.
6. **Hate & Offensive Content Detection** → phát hiện nội dung thù ghét, cực đoan, xúc phạm.

Chú ý: ENABLE -> cho phép

+ Nhấn **ENABLE** ở các mục để chọn các tiêu chí bạn muốn phân tích ở video.

Nudity Detection disable

Detect adult content, nudity and suggestive scenes

Sexual activity
Actual or simulated sexual activity with exposed nudity
`nudity:sexual_activity` ACCEPT REJECT

Sexual display
Explicit exposure of genitals / sexual organs, but not involved in sexual activity
`nudity:sexual_display` ACCEPT REJECT

Erotica
Exposure of breasts, nude buttocks or the pubic region but not of sexual organs
`nudity:erotica` ACCEPT REJECT
Balanced (0.5)

Nudity art
Known paintings, sculptures or statues with nudity
`nudity:art` ACCEPT REJECT

Sextoys
Displays of sextoys that are not in use, such as dildos, plugs, sex dolls or fleshlights
`nudity:sextoy` ACCEPT REJECT

Female underwear
Women wearing lingerie, wearing visible bras (the cup has to be visible, and sports bras are excluded), or wearing visible underwear, panties or thongs
`nudity:female_underwear` ACCEPT REJECT

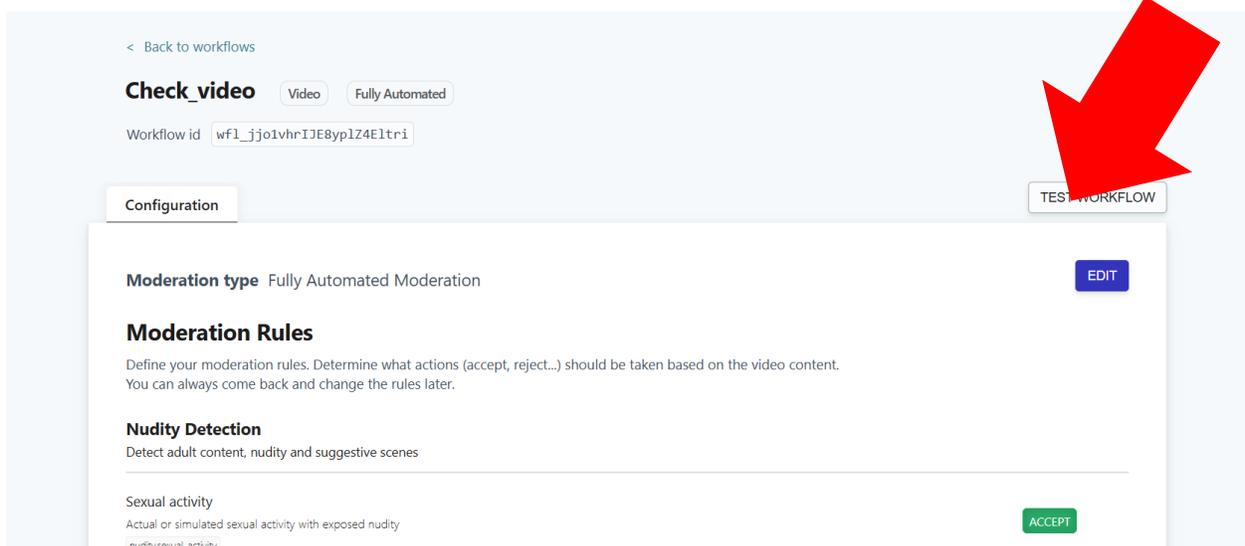
+ Cuối cùng nhấn **SAVE WORKFLOW** để bắt đầu vào phân tích video.

Violence ENABLE

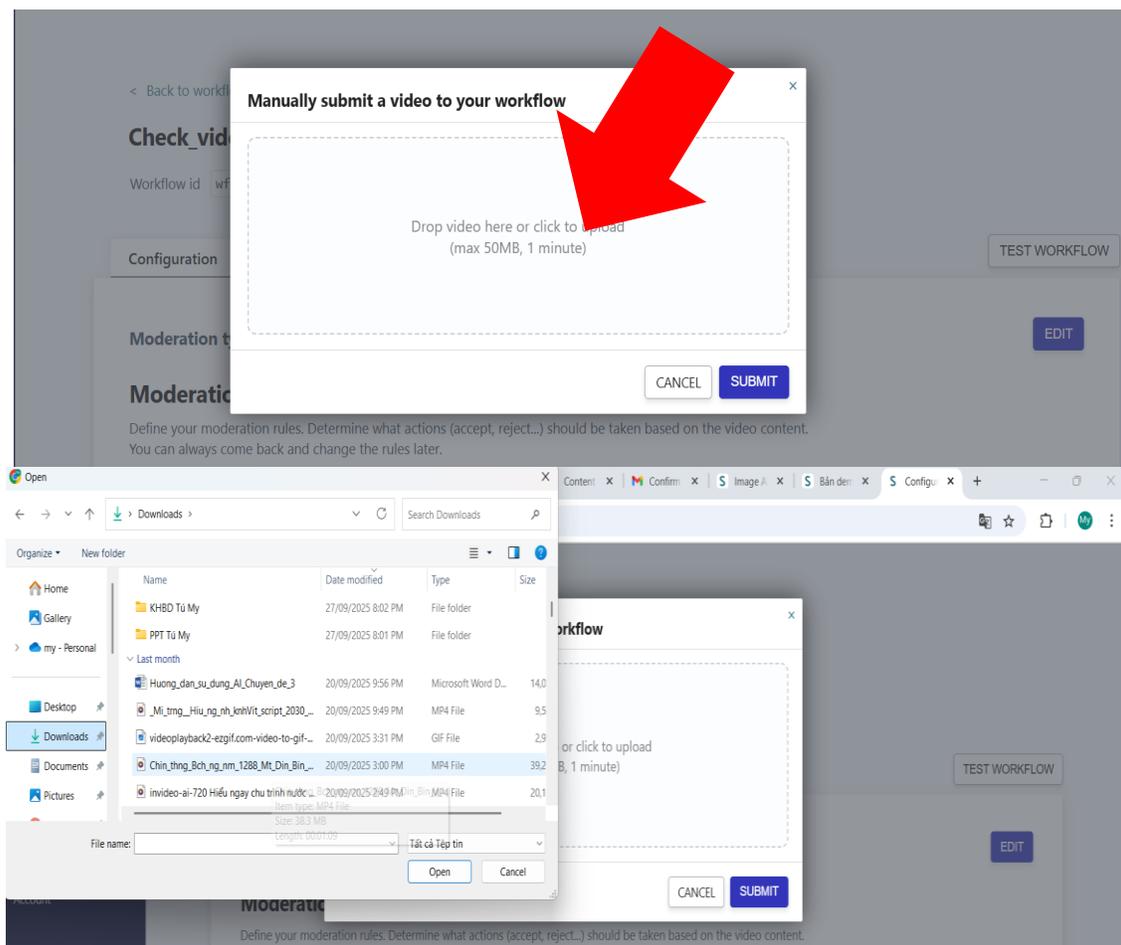
Detect violent scenes

SAVE WORKFLOW

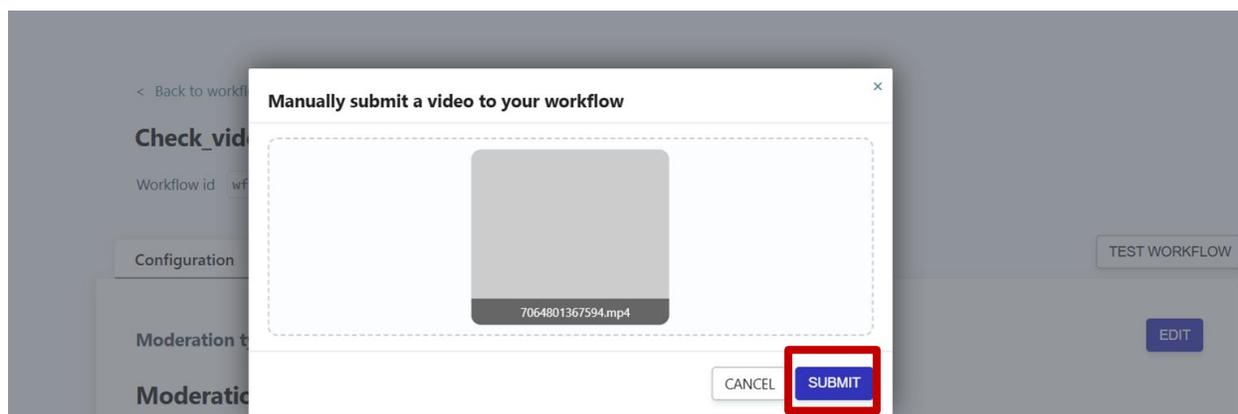
+ Hệ thống sẽ hiện ra giao diện như sau:



-> Nhấn chọn **TEST WOKFLOW** và chọn video bạn muốn.



+ Nhấn **SUBMIT** để hoàn thành tải lên video.



Bước 5: Xem kết quả hệ thống phân tích:

- Với ảnh: hệ thống sẽ hiển thị **kết quả kiểm duyệt** ở khung bên phải

Khảo thân	Các lớp cường độ ?	Các lớp học gợi ý ?
Nội dung dành cho người lớn và gợi ý	Hoạt động tình dục 0%	Bikini 0%
Bạo lực, Máu me, Vũ khí	Hiện thị tình dục 0%	Sự phân cắt 0%
Máu me, bạo lực, tự làm hại bản thân, vũ khí	Khiêu dâm 0%	Đồ lót 0%
Chủ đề thù hận và nhạy cảm	Rất gợi ý 0%	Ngực nam 0%
Nội dung xúc phạm, thù hận hoặc nhạy cảm	Gợi ý 0%	Đồ lót nam 0%
Chất & Tệ nạn	Hơi gợi ý 0%	Minishort 0%
Ma túy, rượu, thuốc lá, cờ bạc	Không có 99%	Váy ngắn 0%
Nội dung văn bản và mã QR	Bối cảnh ?	Nghệ thuật khóa thân 0%
Phân tích văn bản và mã QR	Biển, hồ, hồ bơi 0%	Gợi ý khác 0%
Mô tả nội dung	Ngoài trời khác 42%	Sơ đồ 0%
Chất lượng và loại hình ảnh	Trong nhà khác 57%	Đồ chơi tình dục 0%
		Tập trung gợi ý 0%
		Tư thế gợi cảm 0%
		Đồ bơi nam 0%
		Đồ bơi một mảnh 0%
		Rõ ràng là không mặc quần áo 0%

Chuyển đổi kết quả thô

☞ Nếu % = **0–20%** → an toàn.

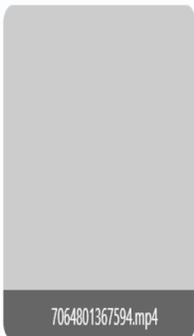
☞ Nếu % = **30–70%** → cần xem xét thêm.

☞ Nếu % = **>70%** → tùy vào nội dung tiêu chí đang xét để đưa ra khuyến nghị, cảnh báo hoặc chặn.

- Với video:

- ↪ **Chấp nhận (xanh lá):** Video an toàn, không chứa cảnh bạo lực/18+.
- ↪ **Từ chối (đỏ):** Video có nội dung vi phạm (bạo lực, máu me, khóa thân...) vượt ngưỡng bạn đã đặt.

Gửi video theo cách thủ công vào quy trình làm việc của bạn



Bản tóm tắt

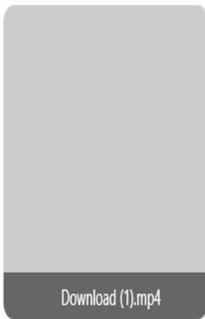
Chấp nhận

Phương tiện được chấp nhận theo quy tắc quy trình làm việc và ngưỡng từ chối của bạn.

Kết quả chi tiết

HỦY BỎ **NỘP**

Gửi video theo cách thủ công vào quy trình làm việc của bạn



Bản tóm tắt

Từ chối 0,98

Phương tiện truyền thông đã từ chối. Quyết định này được đưa ra dựa trên các quy tắc sau: Bạo lực thể xác, Văn bản nhúng, Người nổi tiếng, Súng thật trên tay nhưng không nhắm, Khuôn mặt trẻ em

Kết quả chi tiết

HỦY BỎ **NỘP**