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.

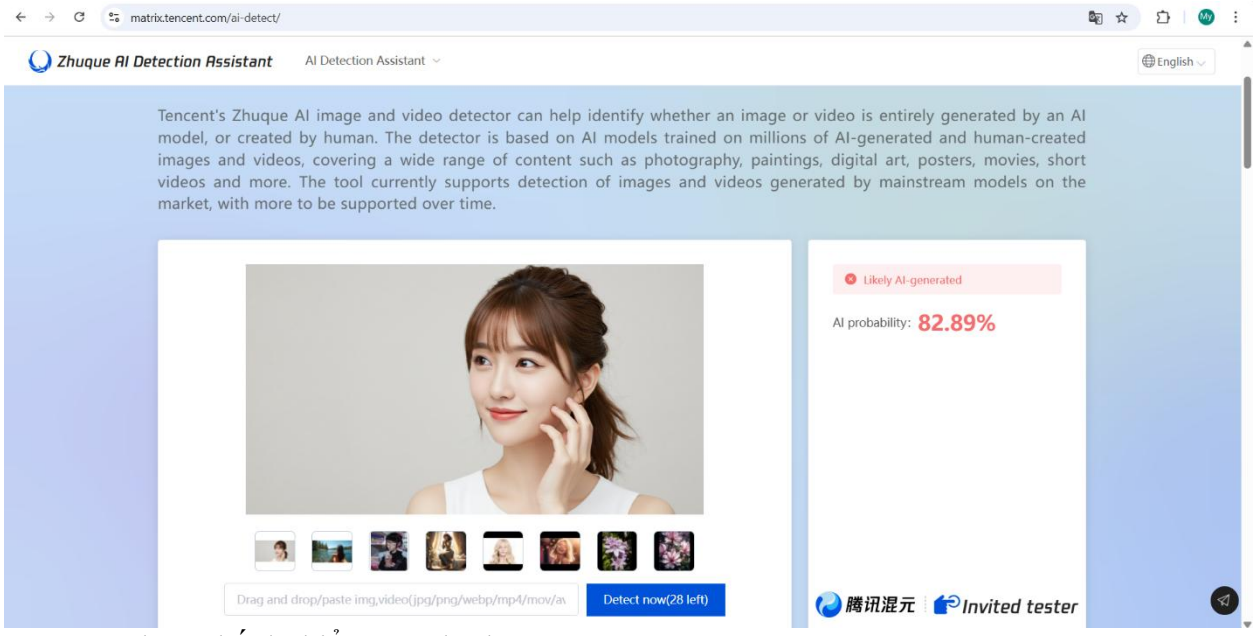
HƯỚNG DẪN SỬ DỤNG AI - CÔNG CỤ PHÁT HIỆN DỮ LIỆU DO AI TẠO RA

Zhuque AI Detection Assistant

Bước 1:

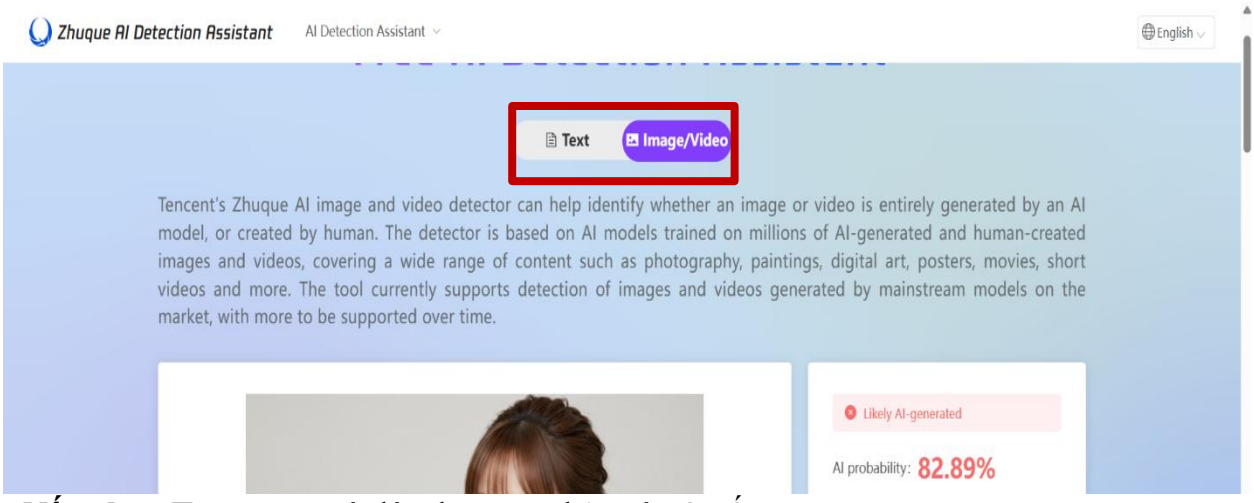
Truy cập trang web: **https://matrix.tencent.com/ai-detect/.**

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



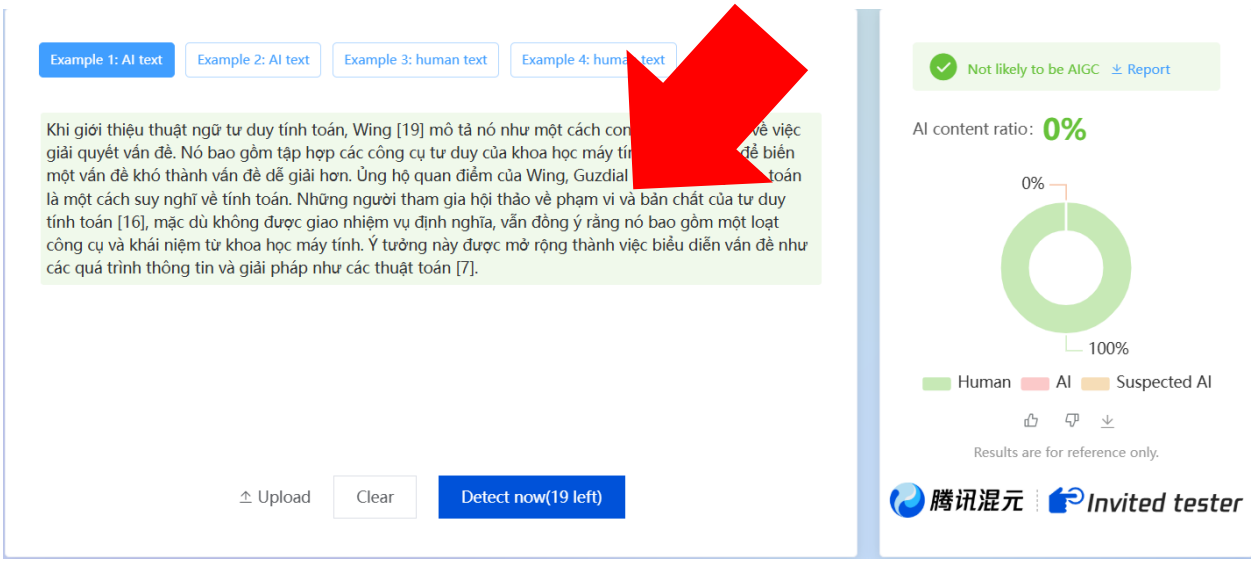
Bước 2: Chọn chế độ kiểm tra phù hợp:

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



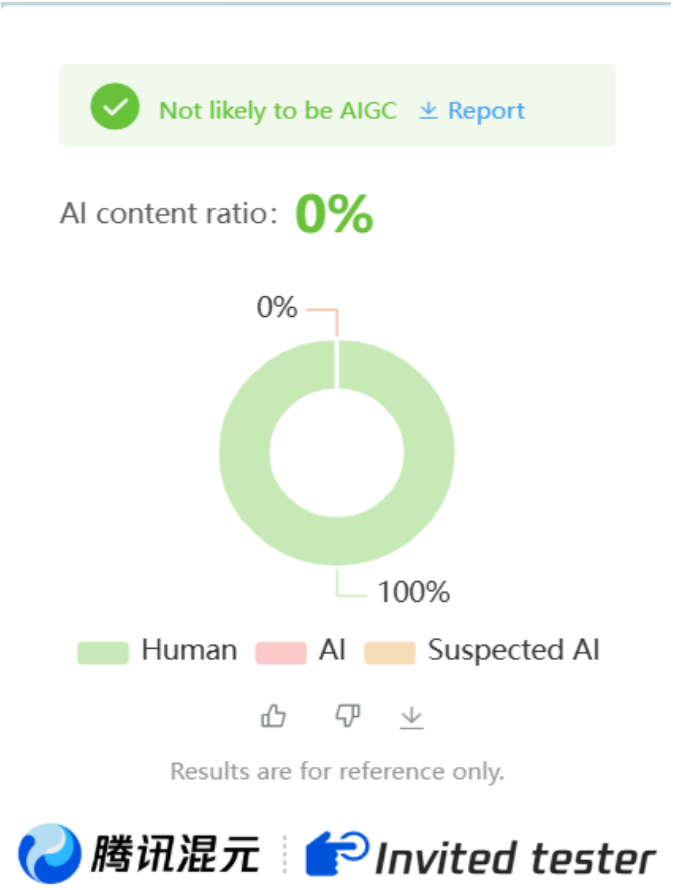
- **Nếu chọn Text**: copy và dán đoạn văn bản vào ô trống.

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

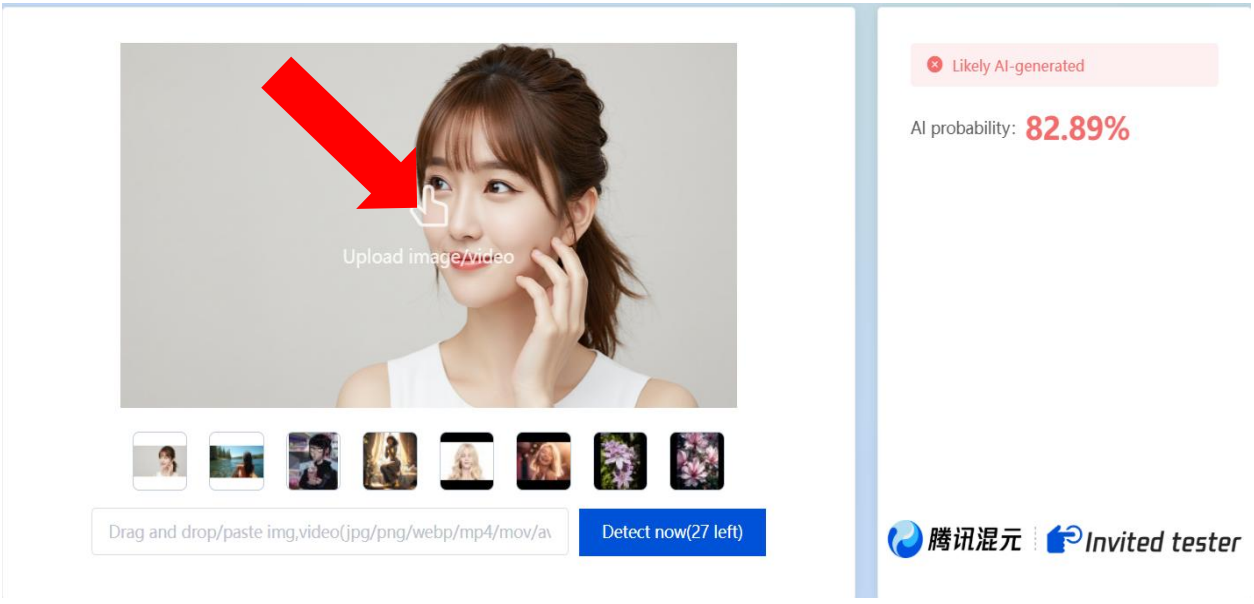


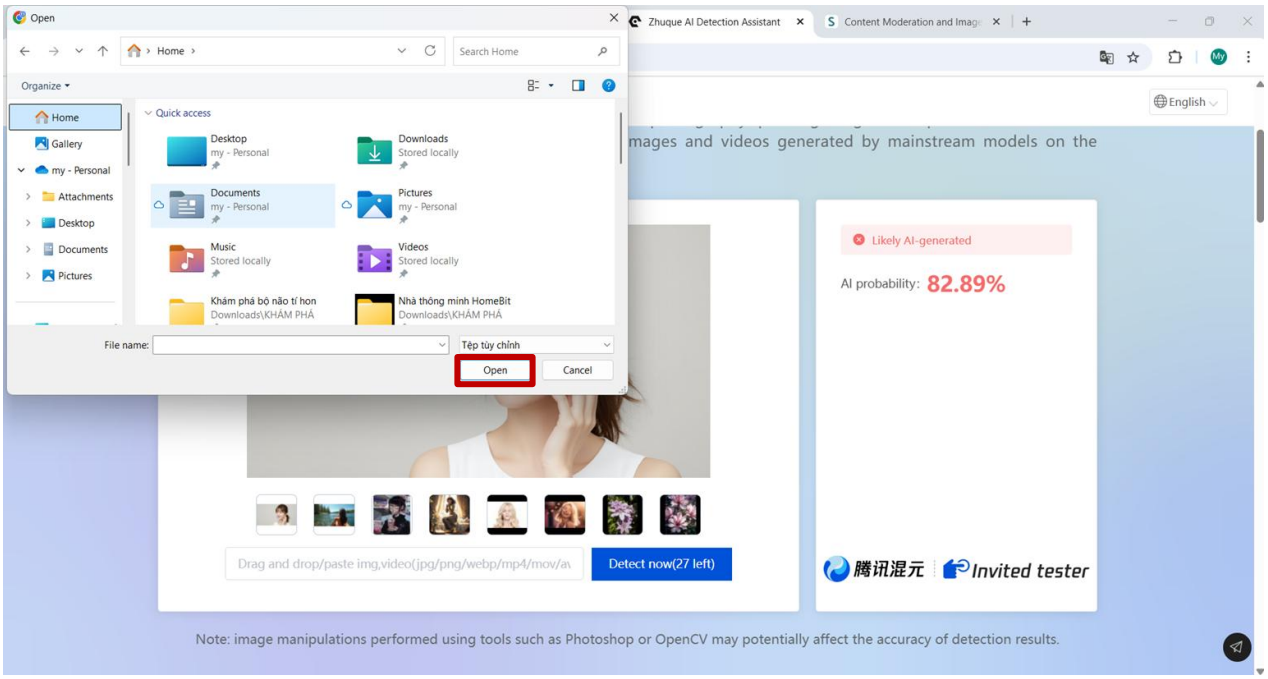
+ Xem kết quả hiển thị ở bên phải màn hình:

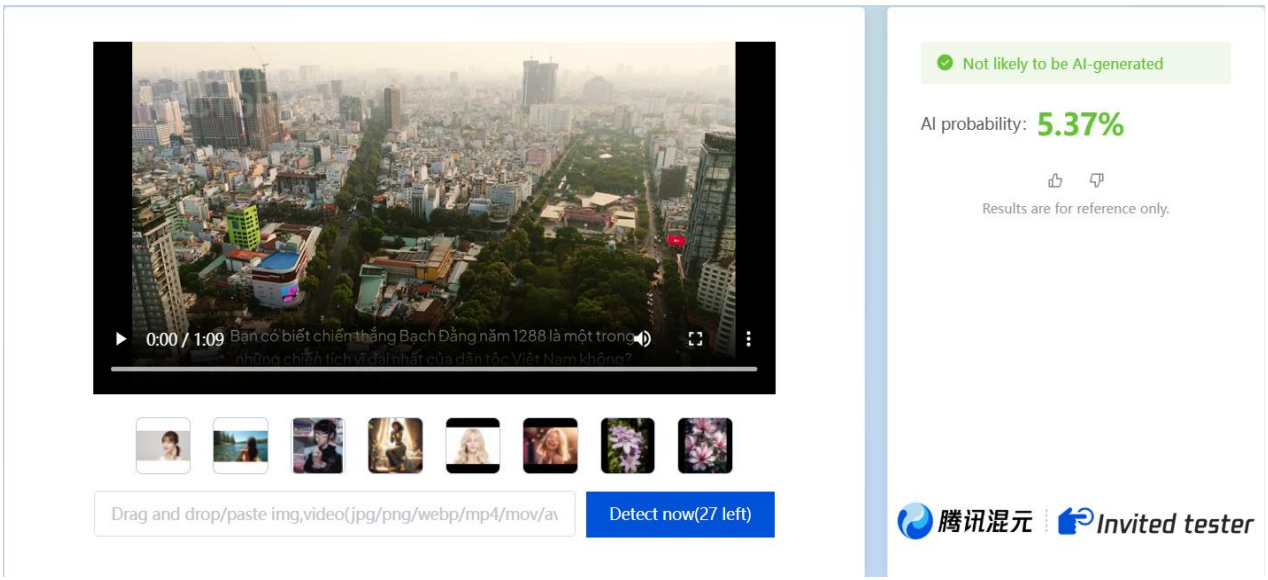
- AI content ratio 0% (màu xanh lá): Nội dung gần như chắc chắn do con người viết.
- Có phần trăm AI (màu đỏ): Cho thấy nội dung có khả năng do AI tạo ra.
- Suspected AI (màu vàng): Hệ thống nghi ngờ có yếu tố AI.



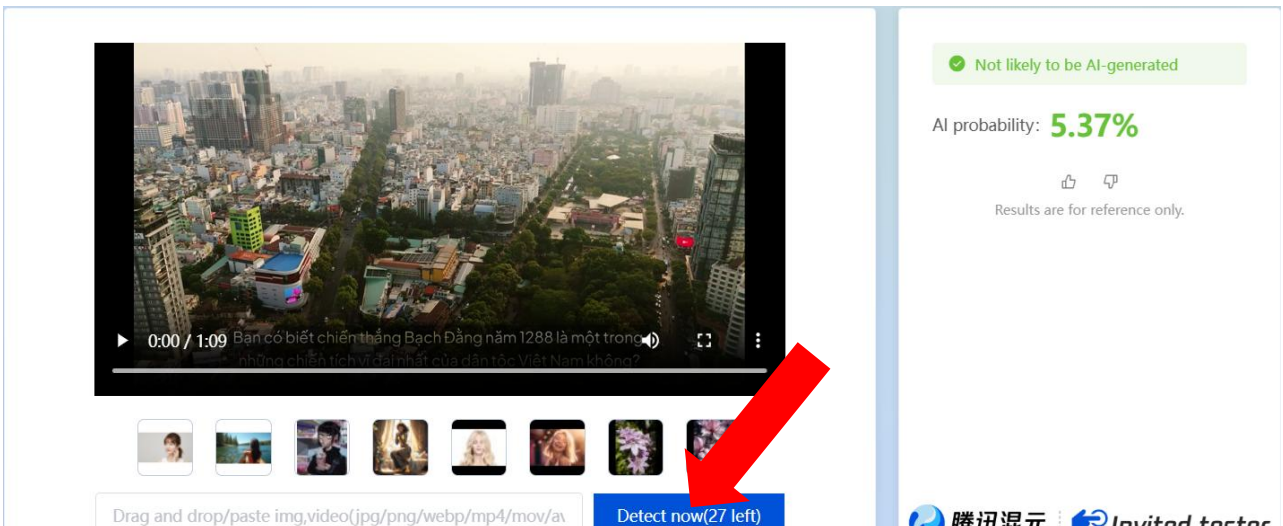
- Nếu chọn Image/Video: tải file ảnh hoặc video lên từ máy tính/điện thoại.







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

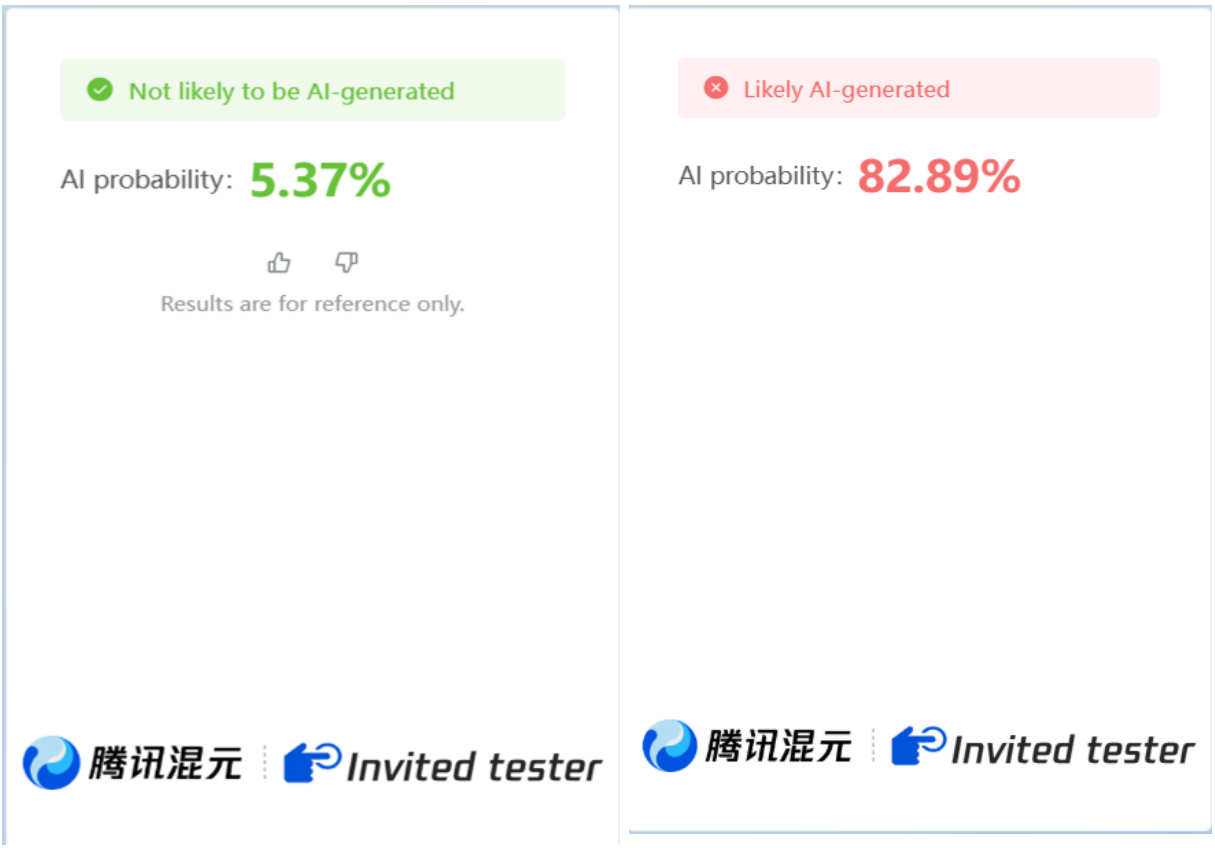


Bước 4: **Xem kết quả trả về:**

- 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.

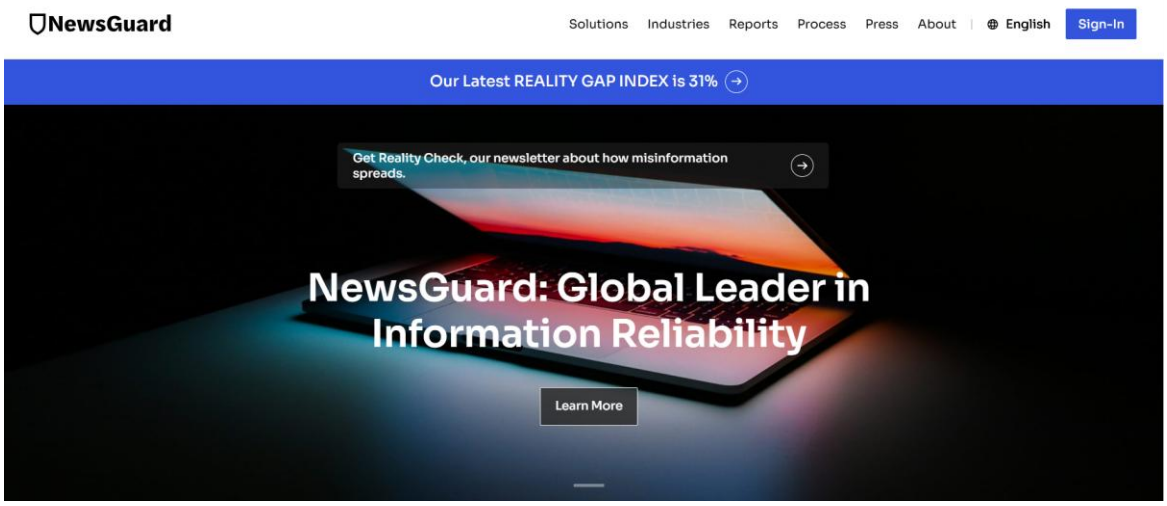


+ *"AI probability: 82.89% – Likely AI-generated"*

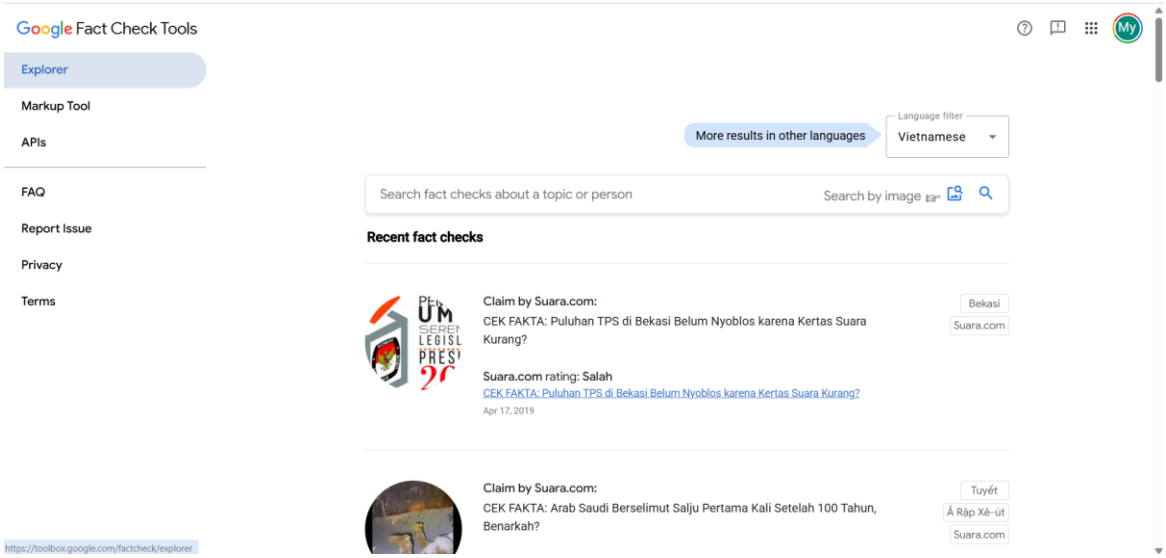
👉 *Nghĩa là xác suất nội dung do AI tạo ra rất cao (trên 80%) → hệ thống cảnh báo đây nhiều khả năng là sản phẩm của AI, cần thận trọng khi sử dụng hoặc chia sẻ.*

- Gợi ý công cụ AI khác

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



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

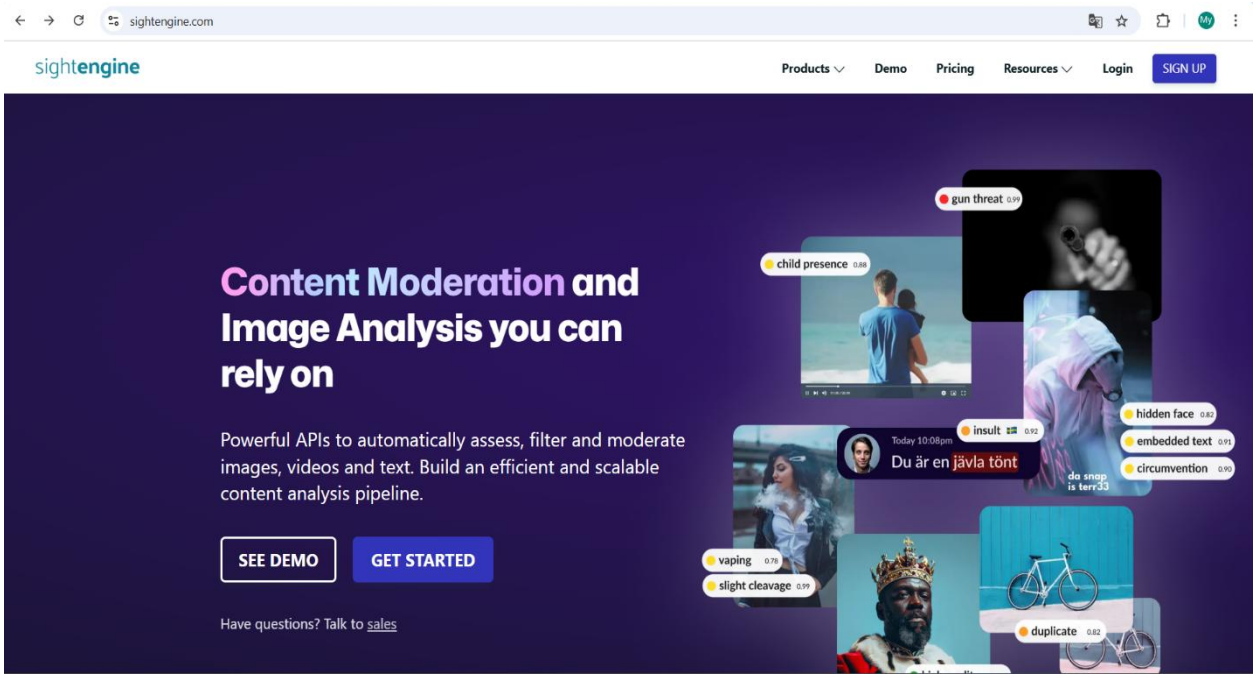


b. Sightengine – Công cụ kiểm duyệt nội dung bạo lực, 18+

- 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.

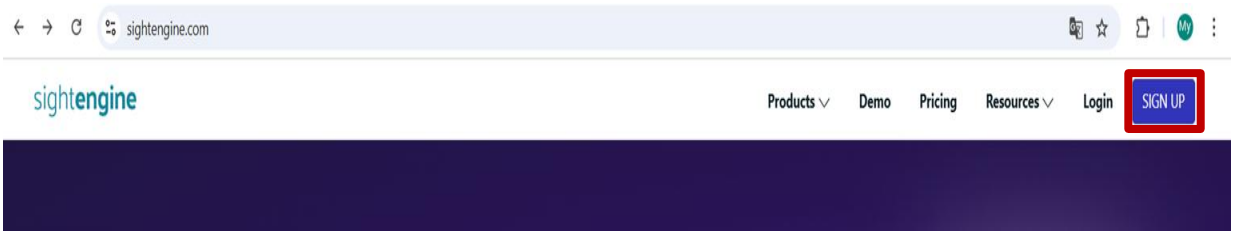
HƯỚNG DẪN SỬ DỤNG SIGHTENGINE – CÔNG CỤ KIỂM DUYỆT NỘI DUNG BẠO LỰC, 18+

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

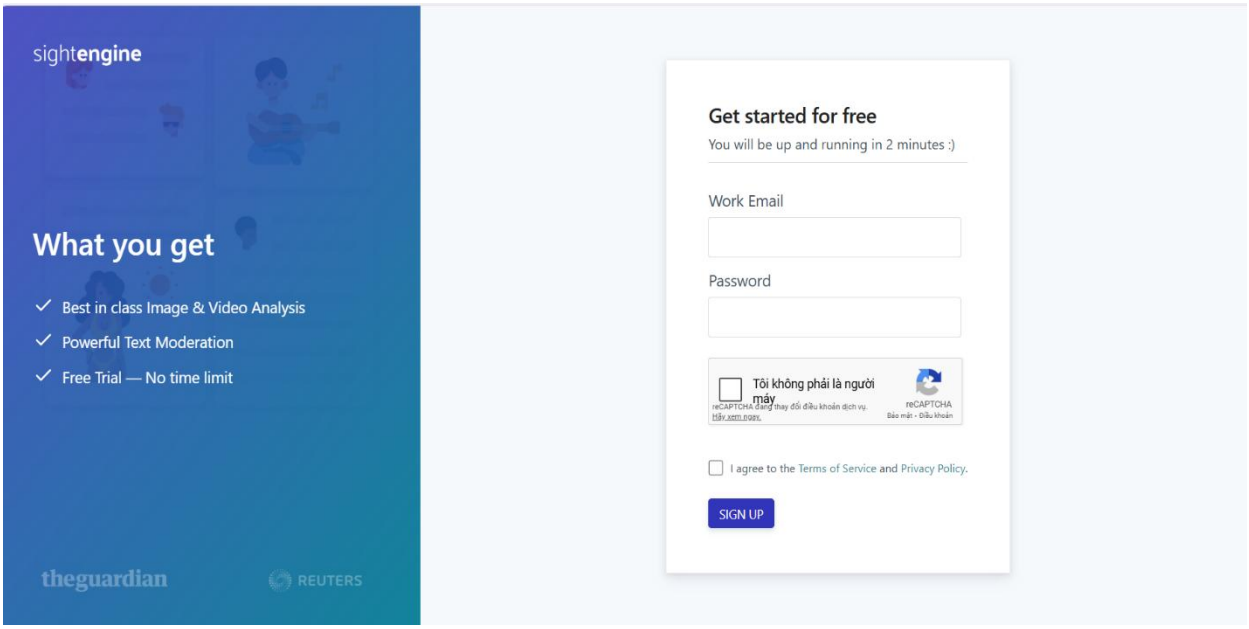


Bước 2: Đăng ký tài khoản miễn phí:

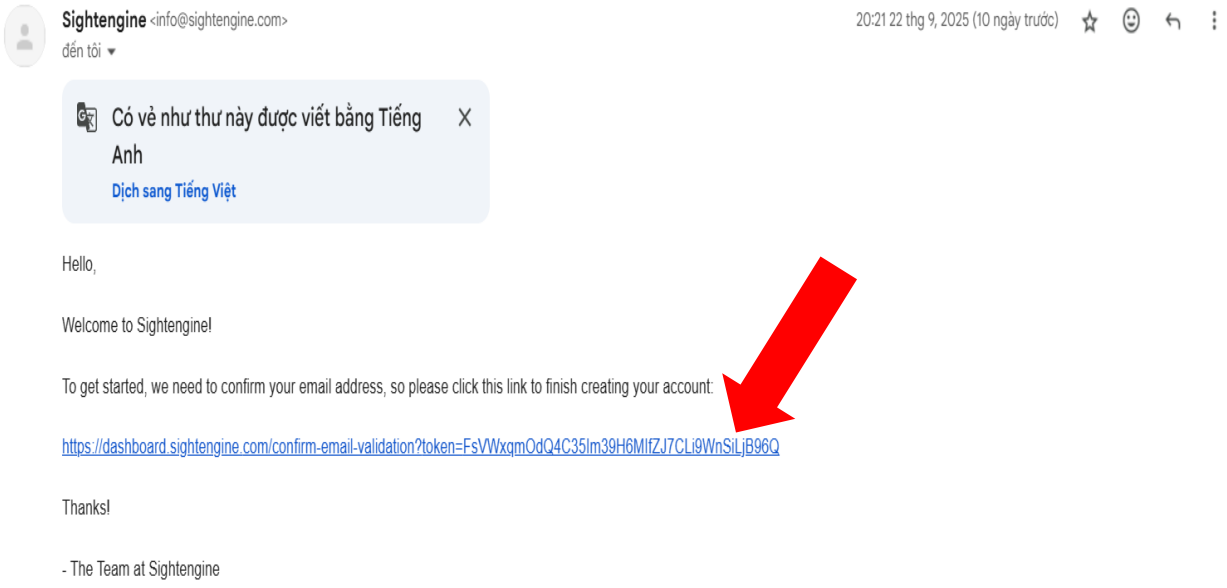
- Nhấn **SIGN UP** ở góc phải.



- 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).

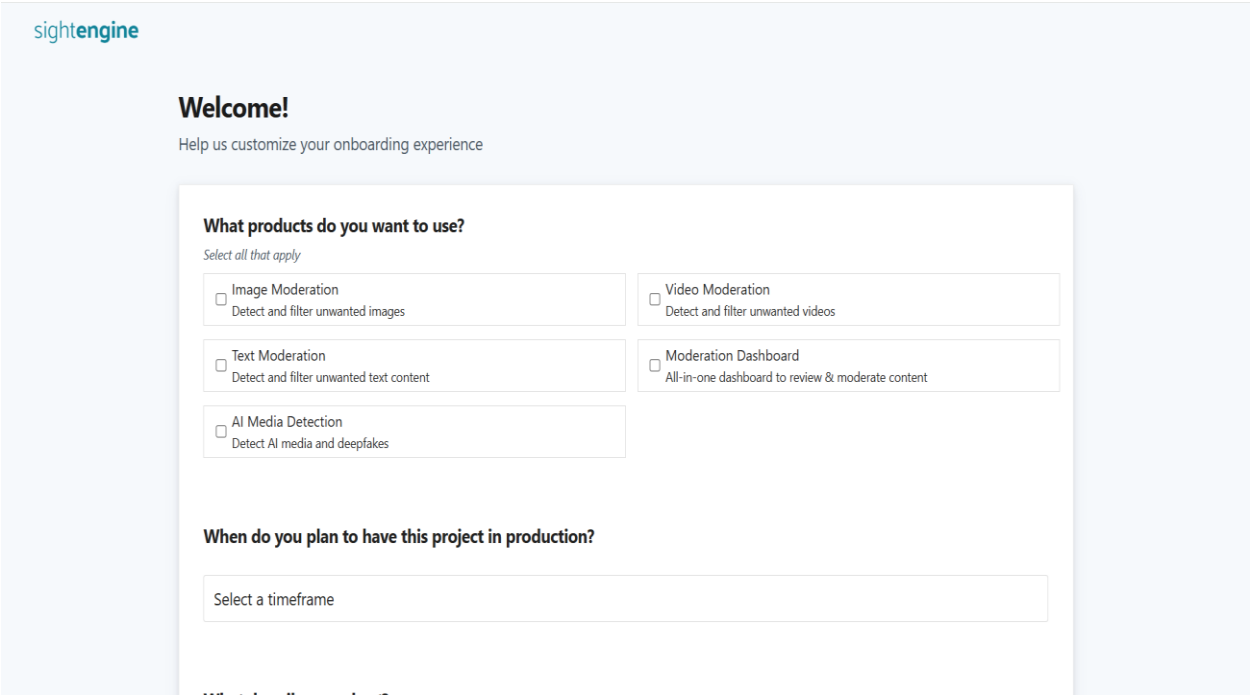
+ **Video Moderation** (lọc video).

+ **Text Moderation** (lọc văn bản).

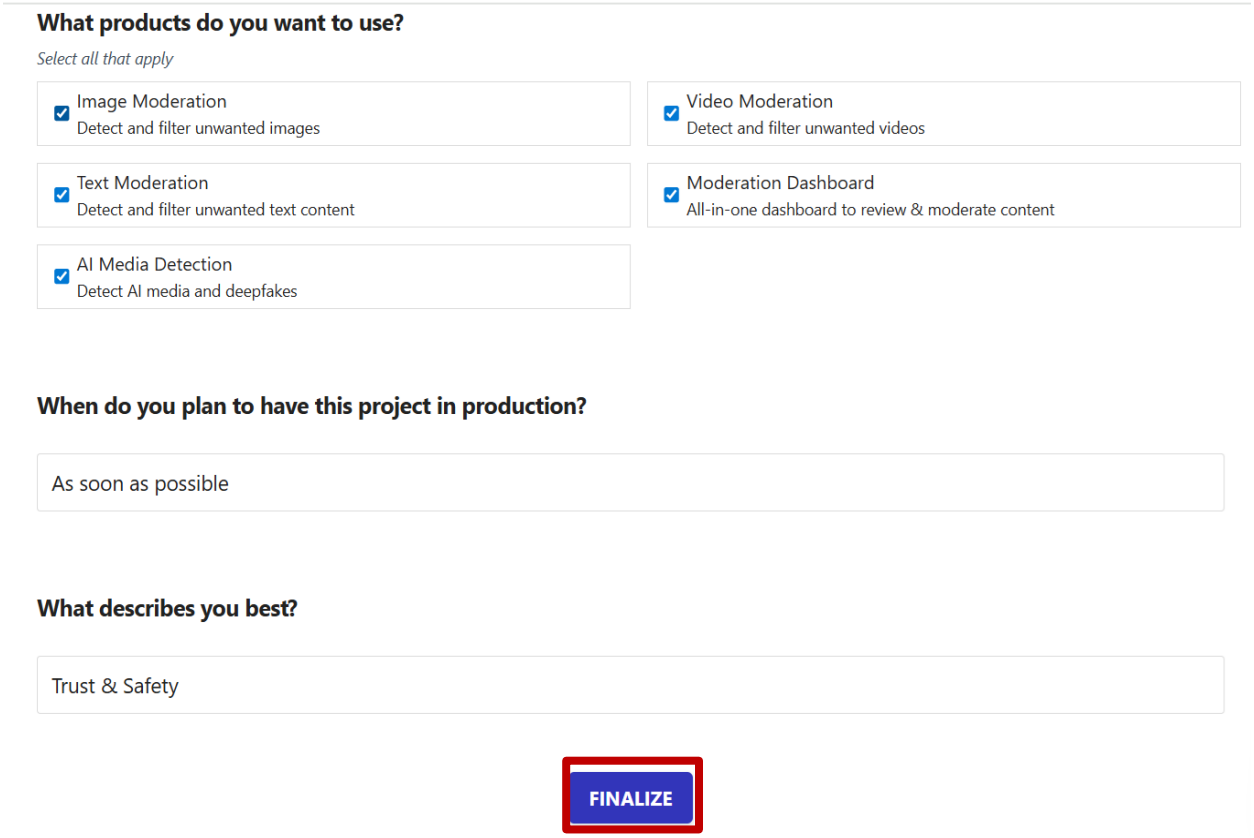
+ **AI Media Detection** (phát hiện nội dung AI, deepfake).

+ **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.

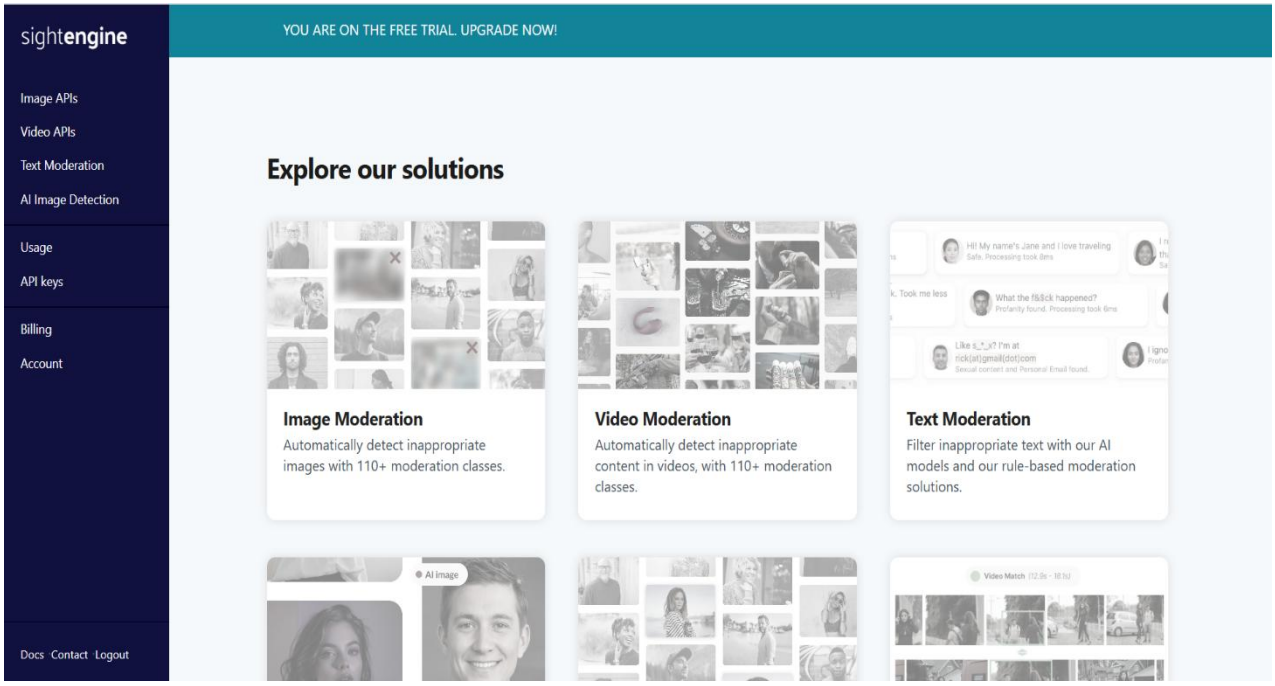


- Cuối dùng nhấn **FINALIZE.**



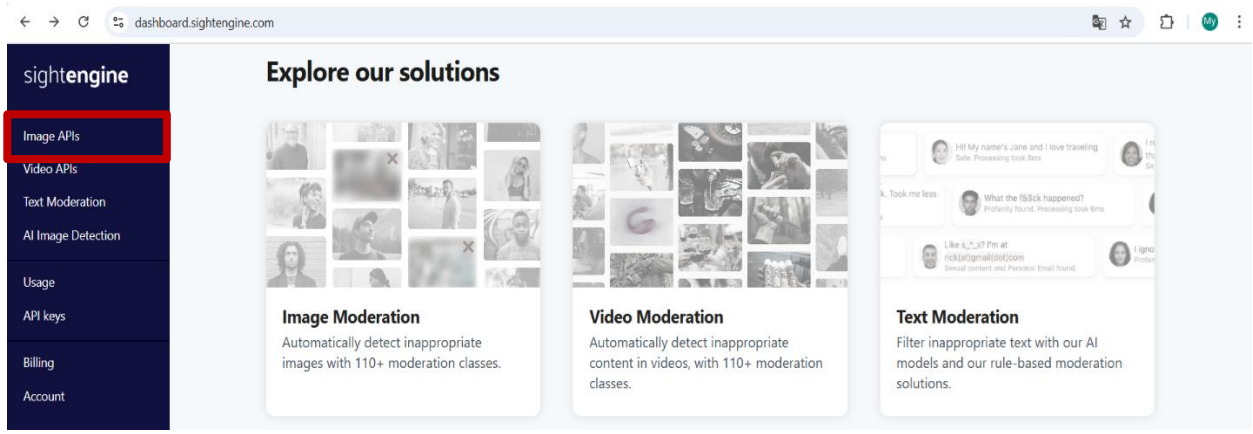
Sau khi đăng nhập, bạn sẽ thấy **Dashboard** với mục **API Keys**.

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

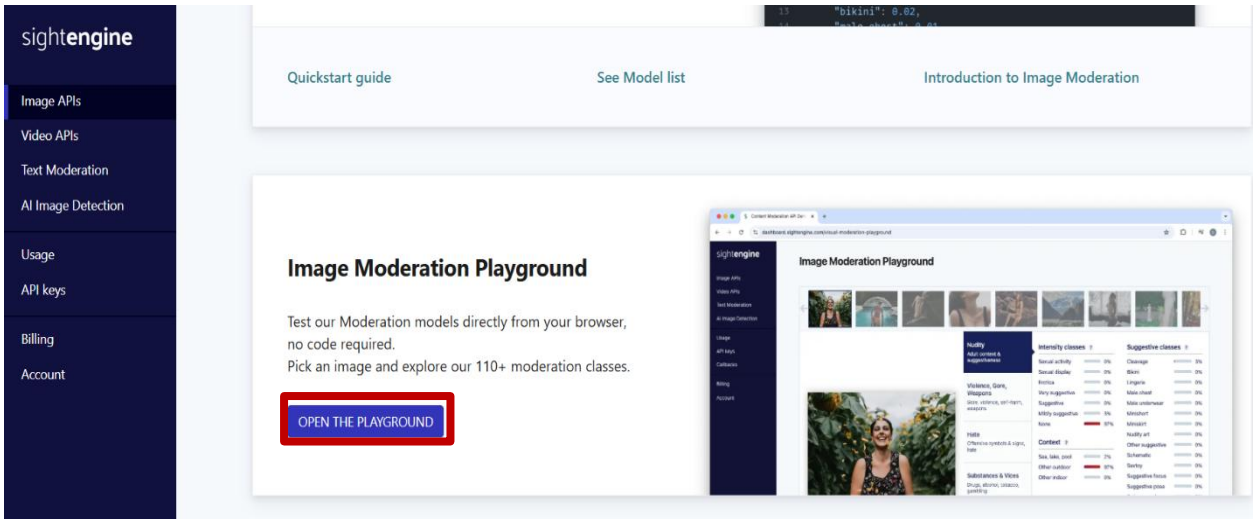


Bước 3: Trong thanh menu bên trái:

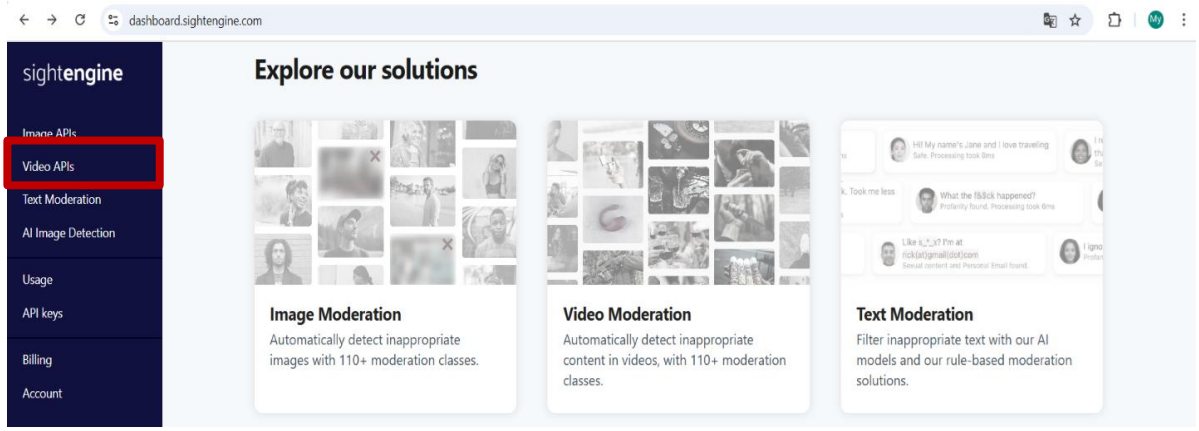
- Chọn **Image APIs** (nếu muốn kiểm tra ảnh).



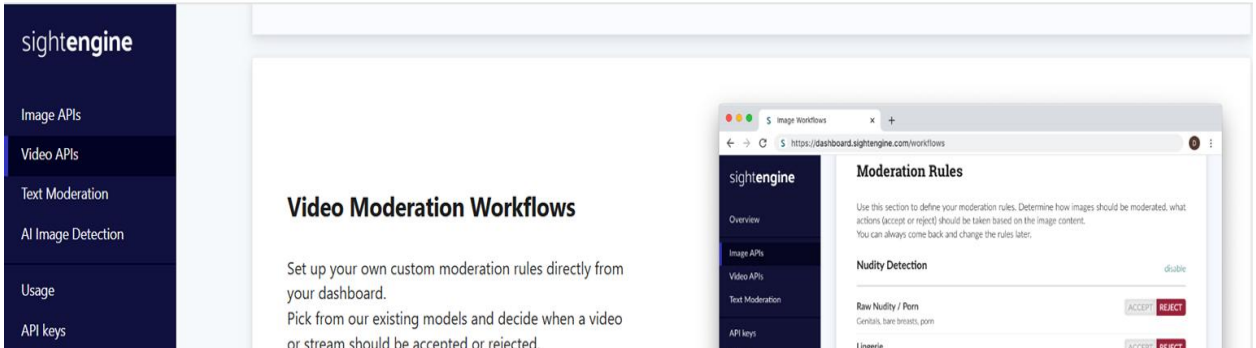
+ Trong Dashboard, tìm mục **Image Moderation Playground** rồi nhấn **OPEN THE PLAYGROUND.**



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

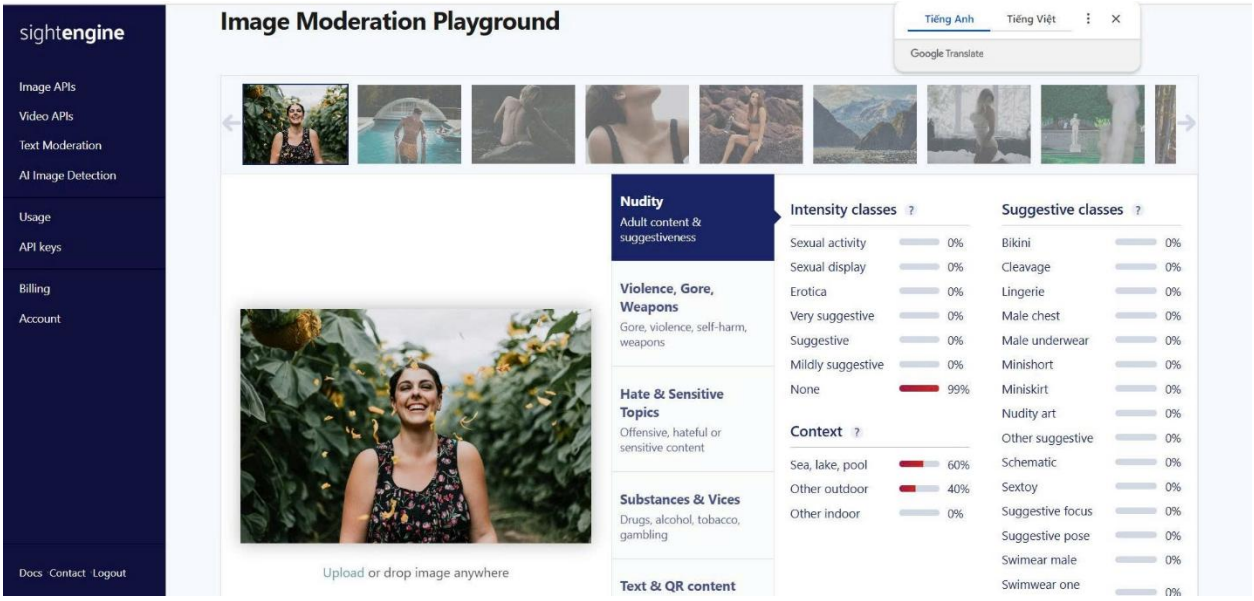


+ Trong Dashboard, tìm mục **Video Moderation Playground** rồi nhấn **GO TO WORKFLOWS.**



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.

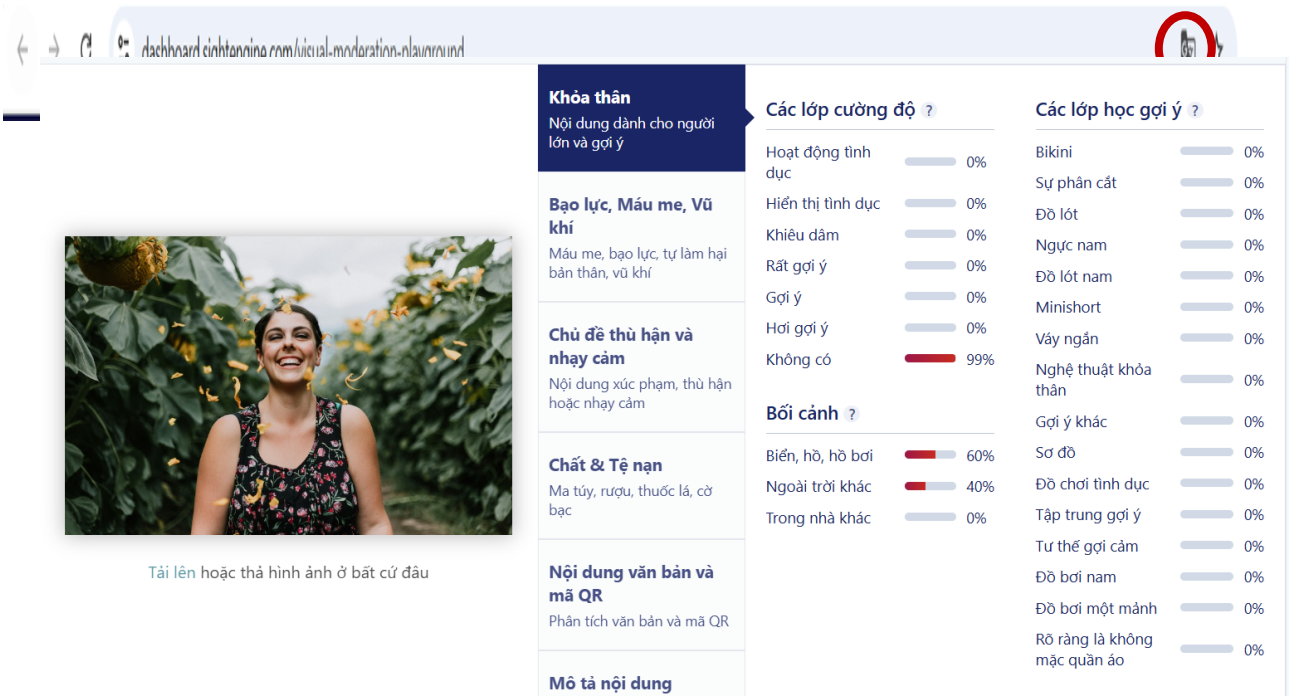


+ Ở giữa màn hình: hiển thị ảnh bạn vừa tải lên.

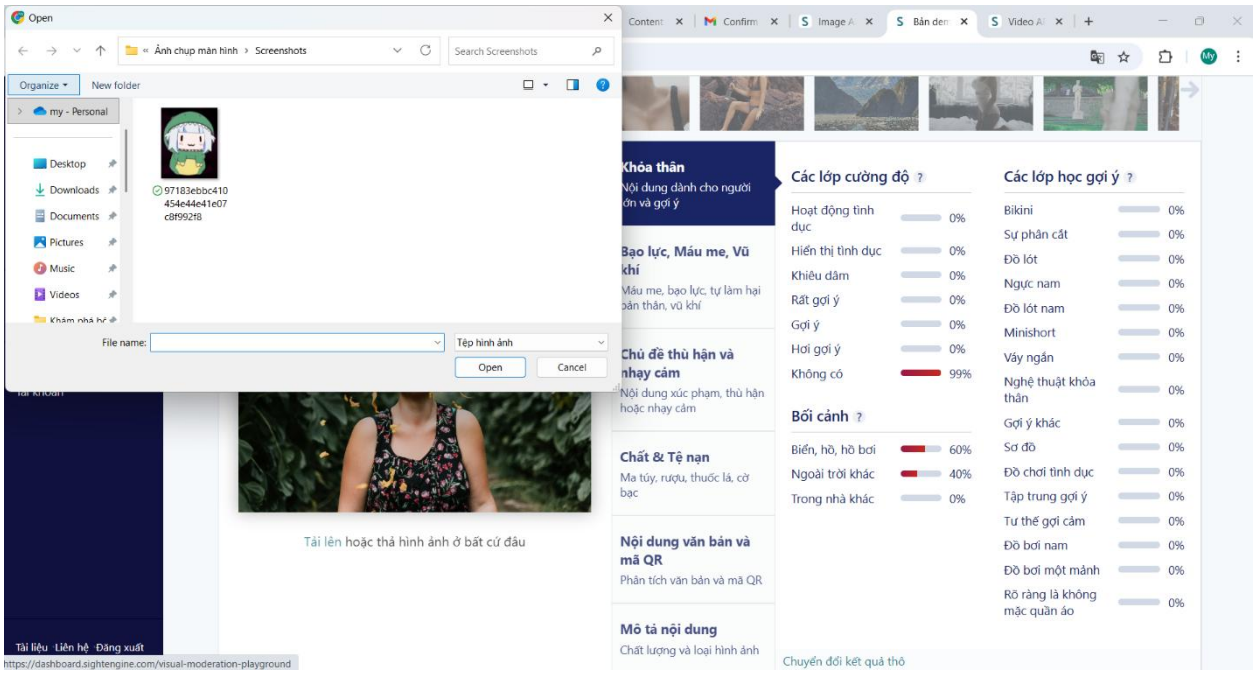
+ Bên phải: là bảng phân tích chi tiết với nhiều nhóm chỉ số.

- **Nudity** (nội dung khỏa 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.

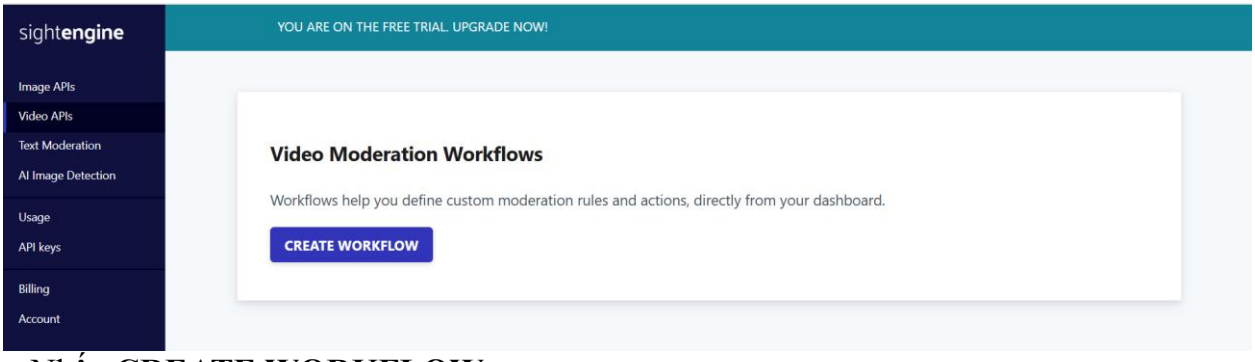


+ Để tải ảnh lên ta nhấn vào chữ **Upload (Tải lên)** và chọn ảnh muốn tải lên từ máy.



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

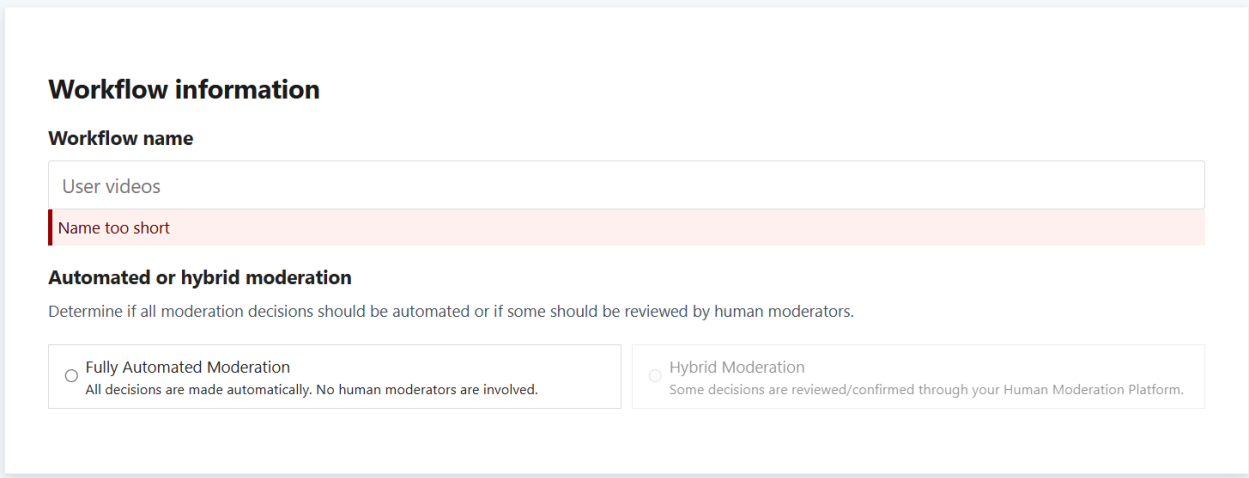
Lưu ý ở đây ta cũng có thể sử dụng Google dịch để hỗ trợ quá trình dịch thuật.



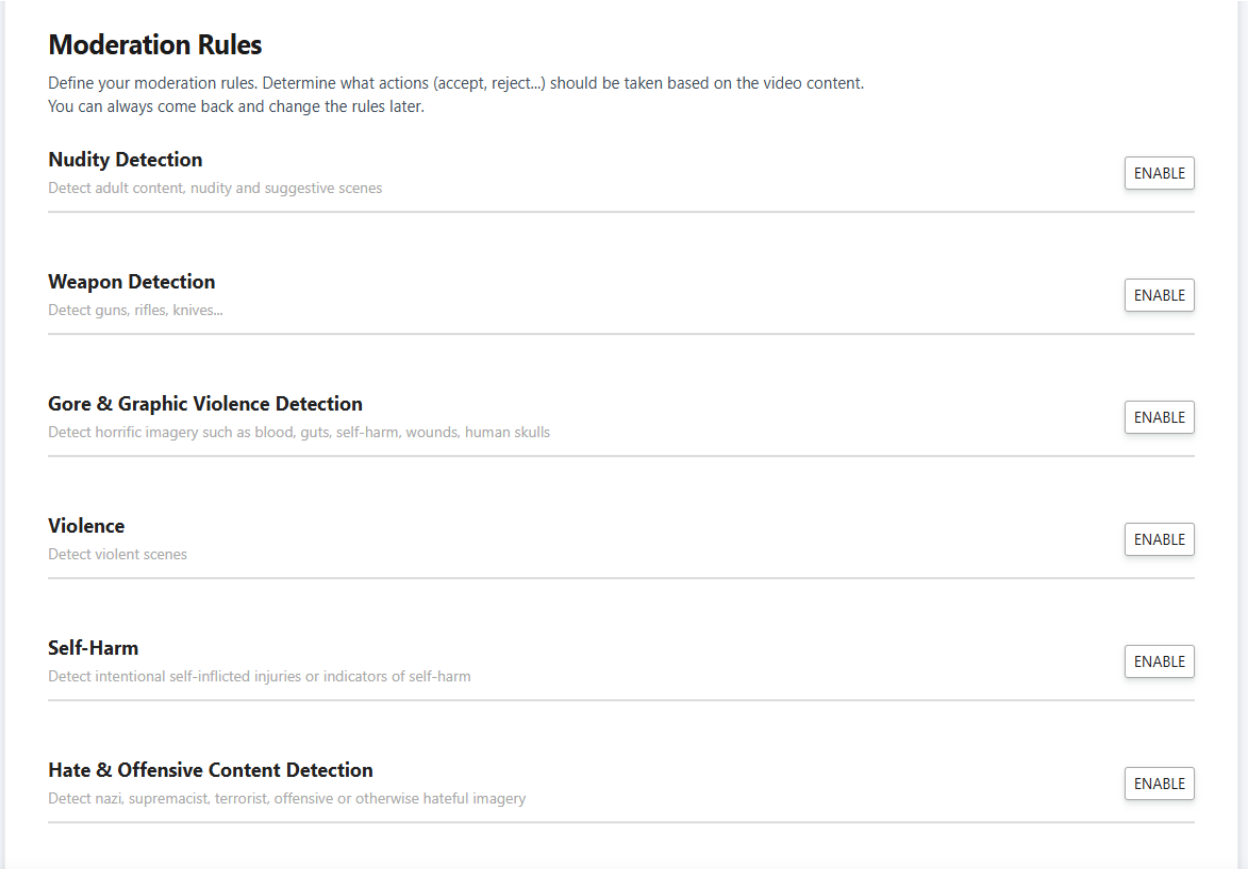
+ Nhấn **CREATE WORKFLOW.**

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 (khuyên dùng để test nhanh).
  - **Hybrid** → cần người duyệt lại một số quyết định.



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

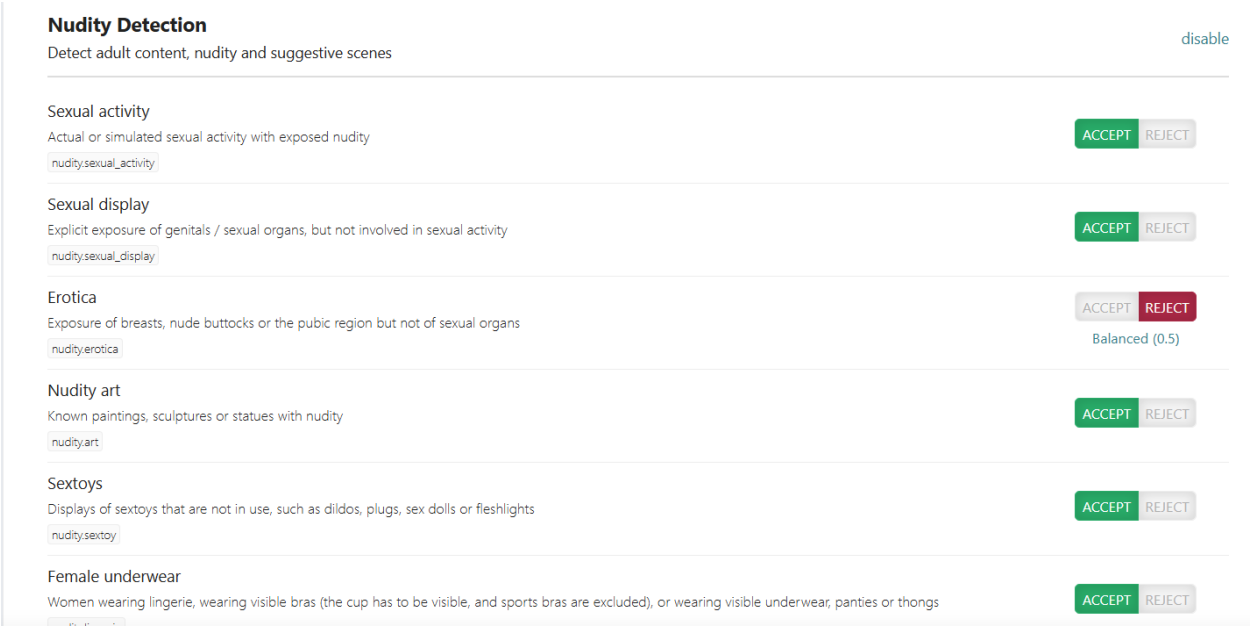


1. **Nudity Detection** → phát hiện nội dung 18+, khỏa thân, cảnh gợi dục.
2. **Weapon Detection** → phát hiện vũ khí (dao, súng, súng trường...).
3. **Gore & Graphic Violence Detection** → phát hiện máu me, vết thương, sọ người...

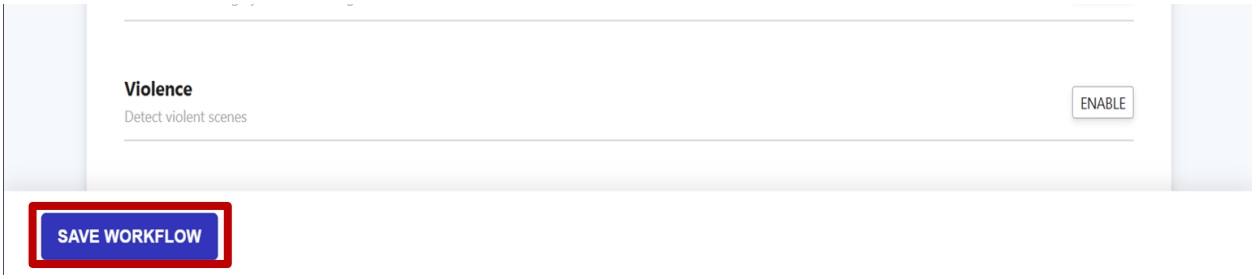
4. **Violence** → phát hiện cảnh bạo lực (đánh nhau, tấn công).
5. **Self-Harm** → phát hiện hành vi tự gây thương tích, tự hại.
6. **Hate & Offensive Content Detection** → phát hiện nội dung thù ghét, cực đoan, xúc phạm.

Chú ý: ENABLE -> cho phép

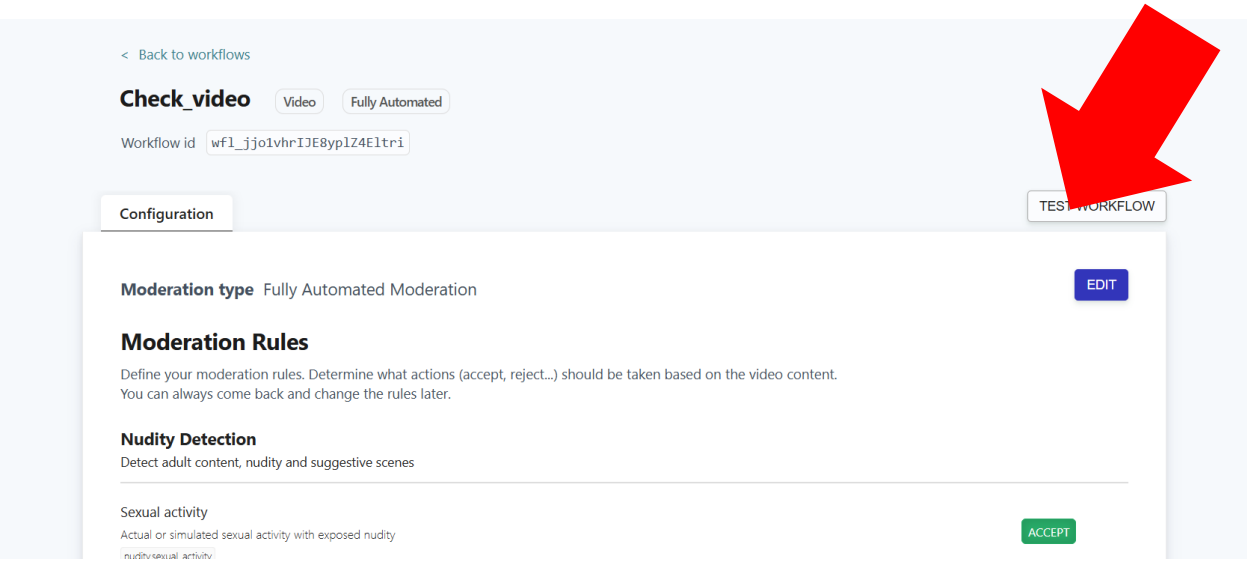
+ Nhấn **ENABLE** ở các mục để chọn các tiêu chí bạn muốn phân tích ở video.



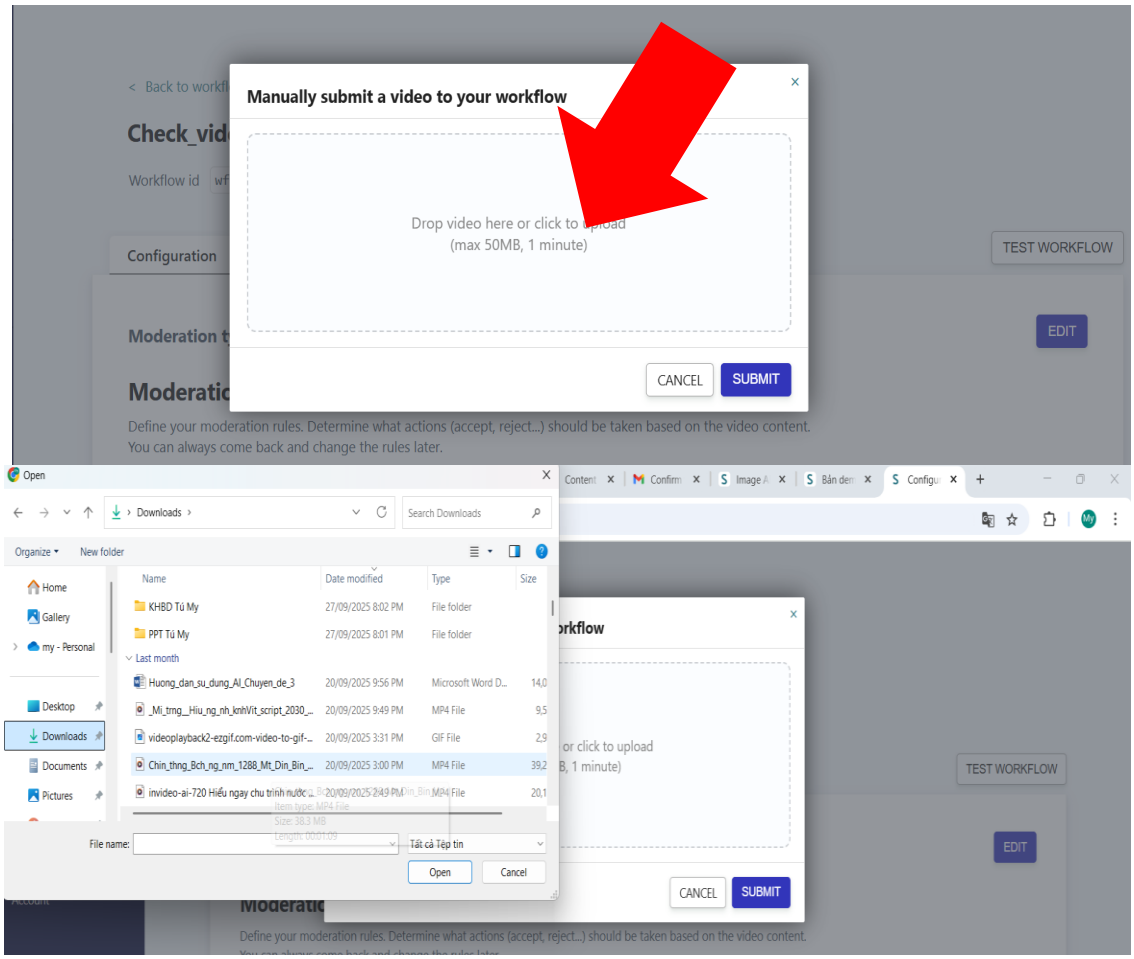
+ Cuối cùng nhấn **SAVE WORKFLOW** để bắt đầu vào phân tích video.



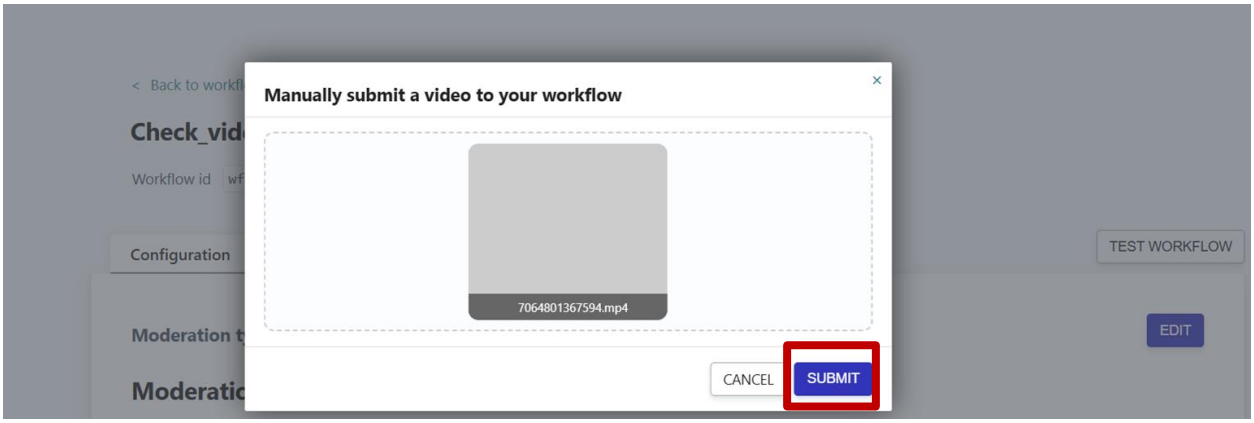
+ Hệ thống sẽ hiện ra giao diện như sau:



-> 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

- 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, khỏa thân…) vượt ngưỡng bạn đã đặt.