

AI-CODE

AI services for COntinuous trust in emerging Digital Enviroments

D4.3

Prototyping the Media Asset Annotation and Management (MAAM) service

Grant Agreement no.	101135437
---------------------	-----------

Project Title	AI-CODE - AI services for COntinuous trust in emerging Digital Environments
---------------	---

Contractual Due date	31/05/2025
----------------------	------------

Actual submission date:	30/05/2025
-------------------------	------------

Lead beneficiary:	CERTH
-------------------	-------



This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101135437

Deliverable number	D4.3
Deliverable Title:	Prototyping the Media Asset Annotation and Management (MAAM) service
Description:	Initial versions of services that were built and prototyped in T4.3.
Type/Nature	☐ R - Document, report ☒ DEM - <i>Demonstrator, pilot, prototype</i> ☐ DEC - <i>Websites, patent filings, videos, etc.</i> ☐ OTHER ☐ ETHICS - <i>Ethics requirement</i> ☐ ORDP - <i>Open Research Data Pilot</i>
Dissemination Level	☒ Public – fully open ☐ Sensitive – limited under the conditions of the Grant Agreement ☐ EU classified – RESTREINT-UE/EU-RESTRICTED, CONFIDENTIEL-UE/EU-CONFIDENTIAL, SECRET-UE/EU-SECRET under Decision 2015/444. For items classified under other rules (e.g. national or international organization), please select the equivalent EU classification level.
Work Package number	WP4
Work Package leader	FBK
Author(s) <i>in alphabetical order</i>	Panagiotis Galopoulos (CERTH) Symeon Papadopoulos (CERTH) Manos Schinas (CERTH)
Internal Reviewers <i>in alphabetical order</i>	Andrea Tagarelli (UNICAL) Miguel Antonio Barbero-Álvarez (UPM)

AUTHORS

Name	Partner	E-Mail
Panagiotis Galopoulos	CERTH	gpan@iti.gr
Symeon Papadopoulos	CERTH	papadop@iti.gr
Manos Schinas	CERTH	manosetro@iti.gr

HISTORY OF CHANGES

Version	Date	Changes
0.1	27/03/2025	Add ToC
0.2	13/05/2025	Drafts of all sections finalized
0.3	21/05/2025	Comments and suggestions by internal reviewers added
0.4	22/05/2025	Feedback from internal reviewers integrated
0.5	27/05/2025	Review by lead beneficiary (CERTH)
0.6	30/05/2025	Review by the coordinator (DSTECH)
1.0	30/05/2025	Final version prepared

TABLE OF CONTENTS

1. Executive summary	5
2. Introduction	5
2.1 Goal and Motivation	6
2.2 Scope of the prototype	6
2.2.1 Scenarios of use addressed by the prototype	7
3. Scientific Rationale	9
4. System Description	10
4.1 Prototype Architecture	10
4.2 Technical Description and Features	13
4.2.1 Asset Annotations with AI Models	17
4.2.2 Visual similarity and near duplicate detection	22
4.2.3 Asset Verification Tools	24
4.3 Model Soundness	27
4.4 Ethical Design Considerations	30
5. Next Steps	34

1. Executive summary

This deliverable presents the architecture and prototype implementation of MAAM (Media Asset Annotation and Management), an AI-powered system designed to support the annotation, organization, and verification of multimedia content. MAAM assists professionals such as journalists, fact-checkers, and disinformation researchers in managing and evaluating the authenticity of images and videos by providing structured, automated assessments through a combination of automatic media tagging, DeepFake detection (DFD), Synthetic Image detection (SID), and forensic tools.

The initial conception and implementation of MAAM was done in the context of the H2020 MediaVerse project¹ and took the form of a decentralized platform for advanced media asset management. In AI-CODE, it has been redesigned and extended as a core service for integrating media verification models and tools. The system is designed to act as a collaborative repository that combines AI-powered analysis with flexible media handling and retrieval capabilities.

This new version of MAAM has been developed in the context of Task 4.3 of Work Package 4 in the AI-CODE project, with a focus on delivering advanced content-oriented services. The platform has reached a “beta stage”, where core functionalities are already implemented and usable, offering substantial support for media annotation and verification tasks. While the system continues to evolve, particularly in areas such as user collaboration and the explainability of AI-based detectors, this prototype represents an early-access version that showcases the platform’s potential and can already serve as a testing ground for a number of AI capabilities in AI-CODE, as a preview of how MAAM can support professionals in identifying manipulated and misleading visual content through integrated, AI-powered verification workflows.

2. Introduction

The prototype is publicly accessible² but requires a two-step user verification process to ensure responsible and controlled use. First, users must register by providing basic personal details along with information about their affiliated organization, followed by email verification. Once this step is completed, the CERTH team reviews the request and manually activates the account after evaluating the user’s profile and intended use of the platform.

The deliverable is structured as follows:

- The remaining parts of **Section 2** explain why MAAM was created and what problems it aims to solve. It also describes what the current prototype includes, how it connects to requirement analysis (D3.1 and D3.2), and the main ways the system can be used.
- In **Section 3**, we describe the ideas and reasoning behind the technical implementation of MAAM. This includes useful input from research and technical advice that helped shape the design and features of the system.

¹ <https://mediaverse-project.eu/>

² <https://maam.mever.gr/app>

- **Section 4** describes how MAAM works in more detail. It explains the main parts of the system, what each component does, and how they work together. It also gives an overview of the key features that are already functional.
- **Section 5** provides an overview of the next steps in the project. It explains what improvements are planned and how MAAM will grow in the future through work in Work Packages 5 and 8.

2.1 Goal and Motivation

MAAM aims to provide a robust, AI-powered platform for the annotation, organization, and verification of multimedia content, designed to support media professionals engaged in disinformation analysis, such as fact-checkers, journalists and social scientists. Originally developed in the H2020 MediaVerse project, MAAM was originally built as a decentralized Digital Asset Management platform, offering core features such as AI-based annotation (e.g., object and action recognition), moderation tagging and content-based retrieval through text-based and visual similarity search, and near-duplicate detection.

The platform has been significantly redesigned and extended within the AI-CODE project to support a broader range of media verification and trust assessment capabilities, while also improving the user experience through a more intuitive and user-friendly interface. These enhancements include the integration of advanced tools for DeepFake detection, synthetic image identification and forensic manipulation analysis. In addition, the system now incorporates more explainable AI outputs by providing detailed model descriptions helping users understand the strengths and limitations of the integrated AI models.

The current prototype implementation specifically addresses the scenario of collaborative media verification, where fact-checkers, journalists, social scientists, and researchers operate within **shared workspaces** to assess the authenticity of digital content. In this context, MAAM now supports several new functionalities introduced in AI-CODE, including:

- Media ingestion from diverse online sources (e.g., media URLs from online platforms such as TikTok)
- User onboarding workflows and access control mechanisms for private, shared, or public workspaces
- Collaboration features for sharing insights for assets
- Explainability and transparency through model-specific information and evidence levels

Together, these features position MAAM as a versatile platform for both managing and verifying multimedia content, enabling users not only to organize and analyze assets but also to assess their credibility in a transparent and collaborative manner. This aligns closely with AI-CODE's broader mission to develop content-driven services for combatting disinformation.

2.2 Scope of the prototype

The Media Asset Annotation Management (MAAM) platform is a web-based tool designed to support media professionals in verifying multimedia content. A detailed overview of MAAM's

key functionalities is provided below.

MAAM serves a dual purpose:

1. **Media Management:** MAAM allows users to efficiently organize, annotate, and search large collections of media assets, streamlining their content management workflows.
2. **Media Verification:** The platform incorporates advanced AI-based analysis tools to evaluate the authenticity of images and videos, aiding in the detection of manipulated or deceptive content.

By integrating comprehensive media organization capabilities with state-of-the-art verification technologies, MAAM offers a powerful and versatile solution for fact-checkers, journalists, and media professionals dedicated to combating visual disinformation.

2.2.1 Scenarios of use addressed by the prototype

This section provides a brief overview of the usage scenarios defined in Deliverable D3.1 and further reined in D3.2, and examines how the current MAAM prototype addresses these scenarios. It highlights the functionalities already implemented that support key user needs and identifies the limitations that must be addressed in the next development phases.

Scenario of use #1: Monitoring and Verifying Media Content During Elections

In preparation for upcoming elections, Richard, a researcher at an NGO focused on election integrity, uses the MAAM platform to manage and verify media content. He creates a dedicated workspace and uploads relevant images and videos using the batch import feature. MAAM's AI tools help him detect manipulated or synthetic media, such as deepfakes and altered images.

The current prototype can largely support this scenario as it already provides detectors for synthetic images, deepfake videos and tampered images. In the next development phase, the integration of newer and more advanced detectors will further enhance the platform's effectiveness, offering greater accuracy and better support to users operating in similar verification contexts. Alongside performance improvements, a key focus will be on explainability and transparency, ensuring that detection results are not only accurate but also interpretable by non-experts. This includes the clear presentation of detection outputs, especially in the case of videos, where results have both a frame-level and a temporal dimension.

Scenario of use #2: Large-Scale Analysis of Deepfake Risks in Videos with Public Figures

Marko, a post-doctoral researcher studying media manipulation, uses MAAM to analyze a large dataset of videos featuring public figures, including suspected deepfakes. He applies MAAM's near-duplicate detection to identify video segments reused in different contexts, then runs deepfake detection models on the flagged segments to assess manipulation. Using MAAM's person identification feature, he links detected deepfakes to specific individuals to determine who is most frequently targeted. Marko compiles his results into a detailed report, highlighting DeepFake risks and manipulation patterns involving public

figures.

MAAM already offers robust capabilities for Near-Duplicate Detection, DeepFake analysis, and person identification. A key area for improvement is the integration of these features in a convenient workflow, enabling them to work in conjunction, as well as the provision of support for report generation.

Scenario of use #3: Detecting Video Manipulations with Near Duplicate Detection

Jonathan, an investigative journalist, uses MAAM to verify the authenticity of a suspicious video showing a government official making a controversial statement. Suspecting manipulation, he applies MAAM's Near-Duplicate Detection (NDD) to search his organization's media archive for the original version. The tool quickly finds a near-identical video, allowing Jonathan to compare the two and confirm that the audio in the circulated version was altered.

While Near-Duplicate Detection is already supported in the current prototype, audio analysis capabilities remain limited, and video comparison features are currently missing.

Scenario of use #4: Identifying Out-of-Context Media with Visual Similarity and Reverse Image Search

Kate, an independent journalist, uses MAAM as part of a collaborative workspace focused on disinformation monitoring. When a viral image falsely linking a disability to a recent vaccine begins circulating, she investigates the potential MAAM's Visual Similarity and reverse image search features. These tools help her trace the image to its original source, a news article unrelated to health or vaccines, proving the image is being misused.

MAAM already includes a visual search feature that allows users to find similar images within their workspaces. However, for it to be fully effective in this scenario it requires that the original media that is relevant for the particular case (e.g. the image as it appeared in the original context) has already been ingested to MAAM. Additionally, web-scale reverse image search is supported, but it is currently limited to redirecting searches to Google Images.

Scenario of Use #6: Repository of Problematic Multimedia Content for Verification

An international media watchdog uses MAAM to build a central archive of known disinformation media, including deepfakes and manipulated images and videos. The team uploads problematic assets and uses MAAM's detection tools to generate detailed analyses, documenting visual and audio manipulation cues. When new suspicious media appears, they compare it against the archive using MAAM's retrieval features to identify similarities. This evolving repository becomes a valuable resource for fast and informed verification, strengthening the organization's ability to track and respond to disinformation.

This scenario is fully supported by the current MAAM prototype. Multiple users from an organization or team can join a shared workspace and use it as a centralized repository for previously verified problematic media. As more users onboard and the volume of stored assets grows, scalability of search within workspaces will also become essential. The current decoupled architecture of the platform facilitates this need, as the underlying storage and

indexing layers, such as Elasticsearch, can be deployed in the cloud with replication and horizontal scaling. This ensures high availability and performance, and fault tolerance even under a heavy search workload.

3. Scientific Rationale

The **Media Asset Annotation and Management (MAAM)** platform is developed in response to the growing need for intelligent, collaborative systems that can manage and verify the vast and increasingly complex landscape of digital multimedia content. A recent report states a **245% YoY increase in deepfakes worldwide**. These findings show considerable growth in DeepFakes in certain countries where elections occur in 2024, including the US, India, Indonesia, Mexico, and South Africa.³ As AI generated media such as synthetic images, DeepFake videos, and manipulated visuals, continue to spread, media professionals face mounting challenges in assessing content authenticity.

MAAM addresses these challenges by integrating advanced media management features with state-of-the-art verification tools, including detectors for synthetic content, forensic manipulation analysis, and DeepFake identification. This approach allows users to organize and annotate their media, and to evaluate their credibility in a streamlined, end-to-end workflow. From a scientific standpoint, MAAM's verification models build on recent breakthroughs in computer vision, deep learning, and explainable AI, delivering both high-performance models and interpretable outputs tailored to non-expert users. It addresses a significant gap in current tools: the lack of integrated platforms that combine content organization, semantic search, AI-powered verification, and collaborative workflows. MAAM supports shared workspaces, multi-user annotations, and verification tasks, enabling teams such as fact-checkers, journalists, and researchers to collaboratively manage, analyze, and validate media assets. This collaborative design enhances transparency, reproducibility, and the scalability of verification processes across distributed teams.

MAAM draws inspiration from initiatives like TrueMedia.org⁴, a non-profit, non-partisan organization that aimed to combat political DeepFakes on social media using AI. Launched in 2024, TrueMedia.org provided free tools for detecting manipulated media, aggregating results from multiple AI detectors to assist journalists and fact-checkers.

However, TrueMedia.org ceased operations in January 2025, leaving a noticeable gap in the landscape of accessible and reliable media verification tools. MAAM addresses this gap not only by integrating similar detection capabilities but also by functioning as an active and evolving platform. It is designed to continuously incorporate new detectors and adapt to the rapidly advancing landscape of generative AI, ensuring that users are equipped to identify even the latest forms of synthetic and manipulated media.

In addition to model-based detection, MAAM supports complementary verification techniques, such as metadata analysis, which can further strengthen the authenticity assessment of

³ <https://sumsub.com/newsroom/deepfake-cases-surge-in-countries-holding-2024-elections-sumsub-research-shows/>

⁴ <https://www.truemedia.org/>

multimedia content. Image and video metadata can be a valuable resource for verifying the authenticity of multimedia content as it provides provenance information, such as the creation date, the device used to capture the media, and the encoding format. By analyzing metadata, it is often possible to uncover discrepancies between the apparent content and the hidden technical details of the file. For instance, metadata may indicate that a file was generated or modified using an encoder-decoder architecture typically associated with DeepFake generation. Such findings can strongly suggest that the image or video is not authentic. To examine metadata, various online tools and software applications are also available, including Exif Info⁵, Jimpl⁶, metadata2go⁷, and Brandfolder – Metadata Extractor⁸. These tools extract and display metadata for further inspection but lack any other detection capability.

In recent years, several commercial solutions have also emerged to detect AI-generated content and particularly DeepFakes, with a strong focus on fraud prevention and enterprise security. Tools such as Reality Defender⁹ and Sensity AI¹⁰ offer multimodal detection systems capable of identifying DeepFake videos, synthetic voices, and manipulated images across various formats. These platforms are designed to help organizations combat impersonation threats, financial fraud, and reputational risks by providing automated alerts and integration options within existing security infrastructures. However, while effective for detecting AI impersonations in a corporate or cybersecurity context, such tools are not tailored for the challenges of political disinformation, which often involve fully generated synthetic content, out-of-context imagery, and subtle manipulations. Other free online tools include Illuminarty¹¹, AI or Not¹² and Hive¹³ for deepfake and synthetic image detection, but most of them provide only these functionalities via APIs and no other functionalities for managing media assets. In addition, each of these tools uses its own method to detect deepfakes, which means there may be limitations on the types of manipulations that can be detected, and great variation in the success rates. In contrast, MAAM is specifically designed to address the needs of media professionals, researchers, and fact-checkers by focusing on visual disinformation in political and journalistic domains, combining detection through multiple state-of-the-art detectors with content management and collaborative verification workflows.

4. System Description

4.1 Prototype Architecture

⁵ <https://exifinfo.org/>

⁶ <https://jimpl.com/>

⁷ <https://www.metadata2go.com/view-metadata>

⁸ <https://brandfolder.com/>

⁹ <https://www.realitydefender.com/>

¹⁰ <https://sensity.ai/>

¹¹ <https://app.illuminarty.ai>

¹² <https://www.aiornot.com/#home>

¹³ <https://thehive.ai/demos/visual-moderation>

MAAM's architecture combines a modular, service-oriented design with advanced AI-driven analysis engines to deliver state-of-the-art media management, annotation and verification platform. It leverages scalable infrastructure, asynchronous communication, and extensible model hosting to ensure robust performance while remaining adaptable to future research developments and feature enhancements.

The core component, called DAM (Digital Asset Management), is a Spring Boot¹⁴ application backed by a PostgreSQL database¹⁵. This persistence layer stores the system's data model, including users, assets, metadata, annotations, and workspaces. Additionally, an Elasticsearch¹⁶ instance is used to index media assets, enabling efficient search and filtering capabilities. The actual media content (e.g., images, videos, audio files) is stored in Docker named volumes¹⁷, which facilitates easy management, portability, and backup of the data. The front-end is built with Vue.js¹⁸ and Nuxt¹⁹ and communicates with the Spring Boot backend through exposed RESTful APIs. In particular, we redesigned the user interface in the first year of AI-CODE, transitioning from JavaScript-based React to TypeScript-based Vue.js and Nuxt, and adopted the Primer design system²⁰ in order to improve consistency, usability, and maintainability across the platform. This redesign will facilitate future revisions of the UI, based on user needs and the specific requirements of different user cohorts (for more details on user cohorts see D3.1 and D3.2).

Beyond the DAM component, MAAM includes four independent services that interact with the core component either via REST-like APIs or through the gRPC protocol²¹. These services are the following:

- a) **Media Annotation Service** is built on the NVIDIA Triton Inference Server²², which hosts multiple annotation models for image and video analysis. It provides several gRPC endpoints for accessing these annotation models. Decoupling the media service from MAAM and its generic annotation model enables easy updates of the annotations integrated into MAAM. The media service can be extended with new annotation types or updated versions of the current annotation models, and MAAM can incorporate these enhancements seamlessly and with minimal effort.
- b) **Near-Duplicate Detection (NDD) Service** supports efficient visual similarity search and comparison in multimedia content, and especially in videos.
- c) **DeepFake and Synthetic Image Detection Service** provides several models for detecting DeepFake videos and synthetic images.
- d) **Image Manipulation Analysis Service** offering multiple methods for performing forensic analysis of uploaded images.

¹⁴ <https://spring.io/projects/spring-boot>

¹⁵ <https://www.postgresql.org/>

¹⁶ <https://www.elastic.co/elasticsearch>

¹⁷ <https://docs.docker.com/engine/storage/volumes/>

¹⁸ <https://vuejs.org/>

¹⁹ <https://nuxt.com/>

²⁰ <https://primer.style/>

²¹ <https://grpc.io/>

²² <https://developer.nvidia.com/dynamo>

All services are containerized and deployed in isolated Docker environments to ensure modularity and ease of management. Figure 1 provides a high-level overview of this architecture.

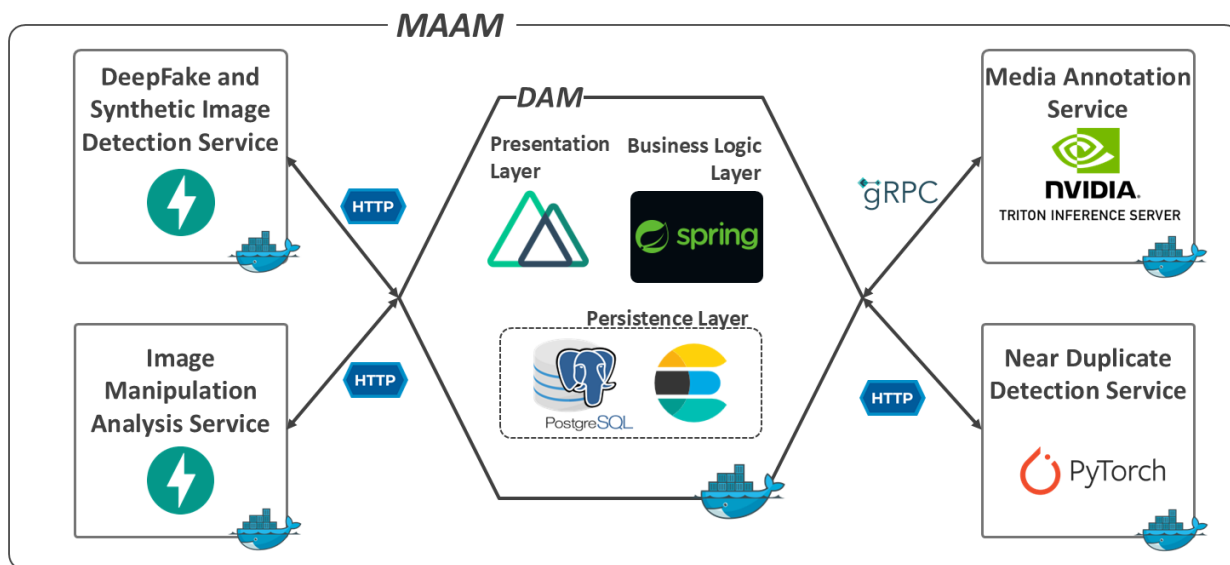


Figure 1 MAAM Service-oriented Architecture

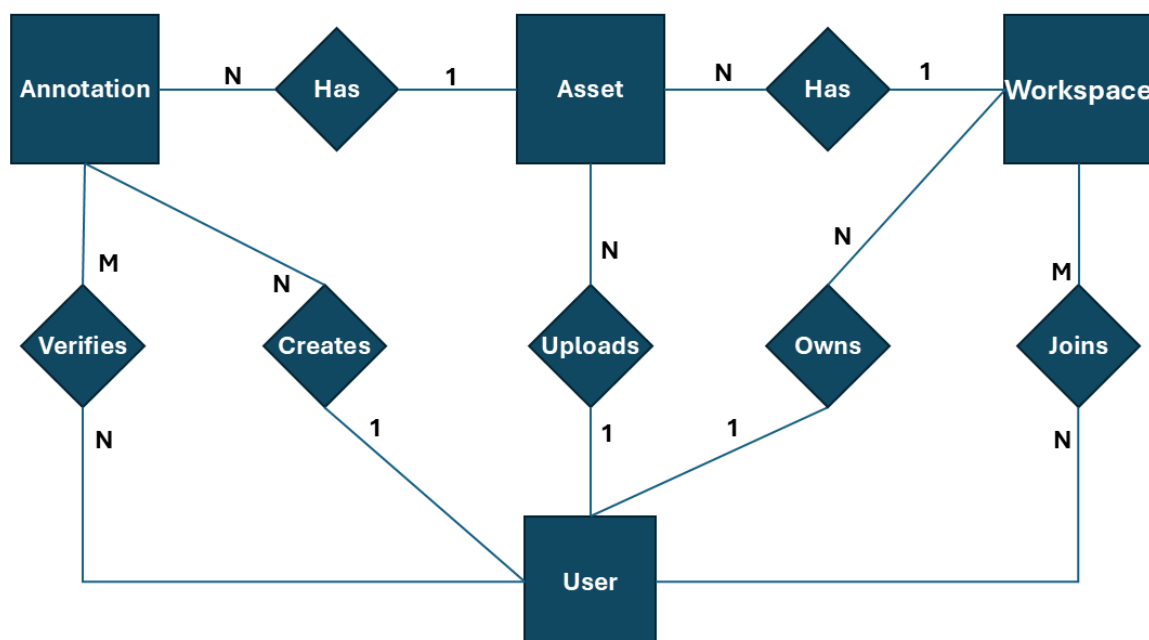


Figure 2 MAAM Data Model

Figure 2 presents an Entity Relationship Diagram (ERD) offering a high-level overview of MAAM's core data model. The diagram highlights the main entities in the system, including

Users, Workspaces, Assets, and Annotations. Each user can create and own multiple workspaces, while also participating in other workspaces as a member. Every workspace can have multiple participants, supporting collaborative workflows. Users can upload assets, but each asset is always associated with a specific workspace. Assets may have multiple annotations, which can vary by annotation type. Annotations can be generated automatically by the system or manually added by users. Users participating in a workspace can create their own annotations but also verify existing ones (produced by the system or by other users) ensuring a collaborative and trustworthy annotation process.

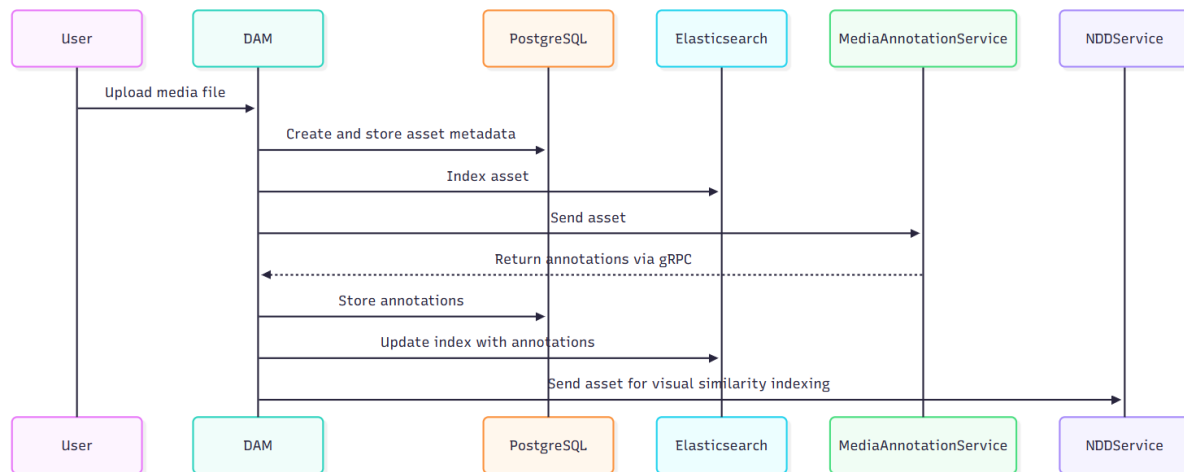


Figure 3 Upload Asset Sequence Diagram

When a user uploads a media file to the DAM (Digital Asset Management) component of MAAM, the platform initiates a process to store, annotate, and index the asset. The uploaded media file is physically stored in a Docker volume, utilizing the local file system of the server where MAAM is deployed. In parallel, the asset metadata such as filename, type, and uploader are stored in a PostgreSQL database. The DAM also indexes the asset in Elasticsearch, enabling fast, filterable search based on user provided metadata. The asset is sent to the Media Annotation Service via gRPC, where AI models generate descriptive annotations (e.g., objects, actions, moderation labels). These annotations are returned to the DAM, are saved to PostgreSQL and used to update the Elasticsearch index, enriching the searchable metadata associated with the asset. Finally, the DAM forwards the asset to the NDD Service, which indexes the visual content for near-duplicate and similarity-based search, allowing the system to detect visually similar content across the repository. This pipeline is presented in Figure 3. In contrast to annotation and NDD indexing, DeepFake and synthetic content detection (for images and videos) and forensic analysis, are performed on demand, initiated explicitly by the end-user through the MAAM interface. Upon request, the DAM sends the selected asset to the appropriate service (e.g., DeepFake detector, synthetic image classifier, or forensic analyzer). The responses are then stored in PostgreSQL and indexed in Elasticsearch, just like the automatic annotations, making the verification results available for search, filtering, and visualization.

4.2 Technical Description and Features

MAAM offers robust media management capabilities, enabling users to store, organize, and

retrieve media assets with ease. These features help users structure their content, annotate important details, and quickly access relevant images and videos for verification tasks.

The platform supports the creation and management of *workspaces*, which act as structured collections of media assets organized around a common theme or investigation. A user can create multiple workspaces or join existing ones and easily switch between them through the interface (Figure 4). Workspaces are designed for collaboration: the workspace owner can invite other MAAM users to join, allowing multiple participants to contribute within a shared environment. More precisely, workspace owners can either invite new users directly within MAAM (Figure 5) or send them an invitation link e.g. via email (Figure 6). In both cases the user must be a valid registered user in MAAM and accept the invitation. This collaborative setup streamlines the process of organizing, analyzing, and verifying media assets as part of a team or coordinated effort.

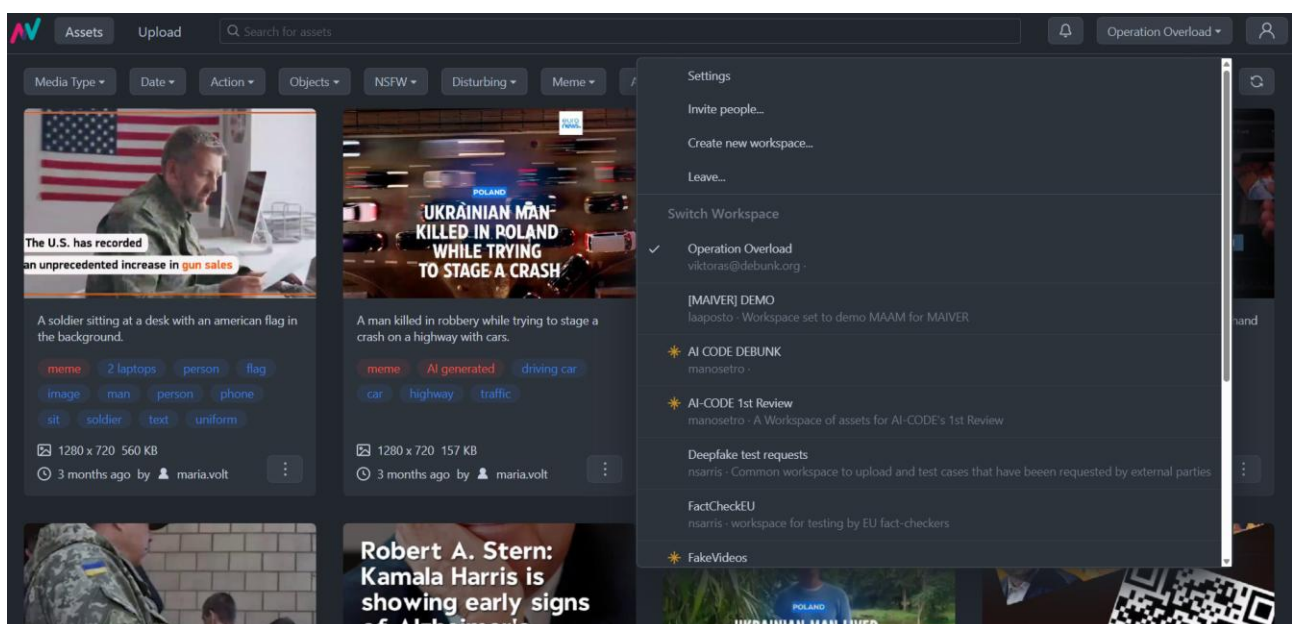


Figure 4 Workspaces in MAAM

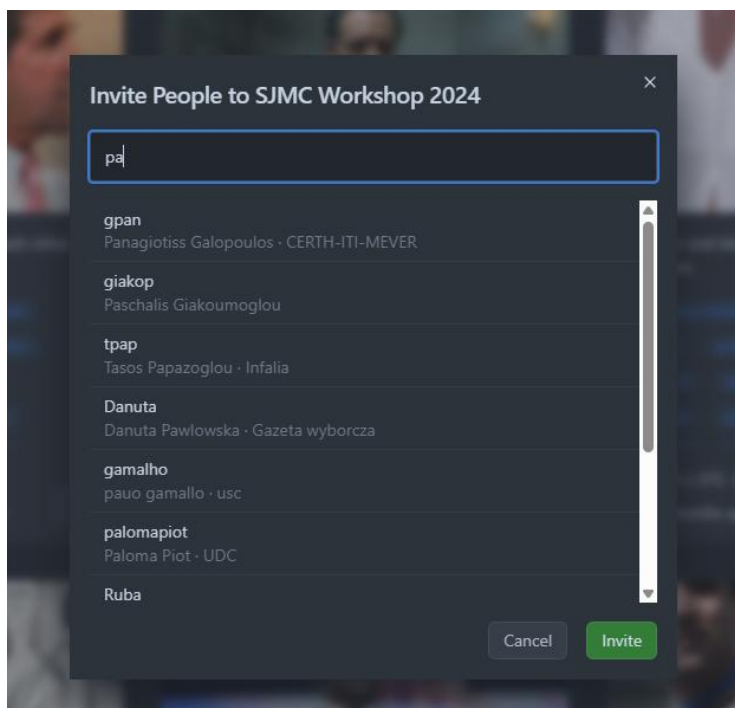


Figure 5 Inviting users to a workspace

SJMC Workshop 2024 Settings Delete Workspace

Properties

Name	Description
SJMC Workshop 2024	Workspace for SJMC Workspace

Invite Link Delete Link

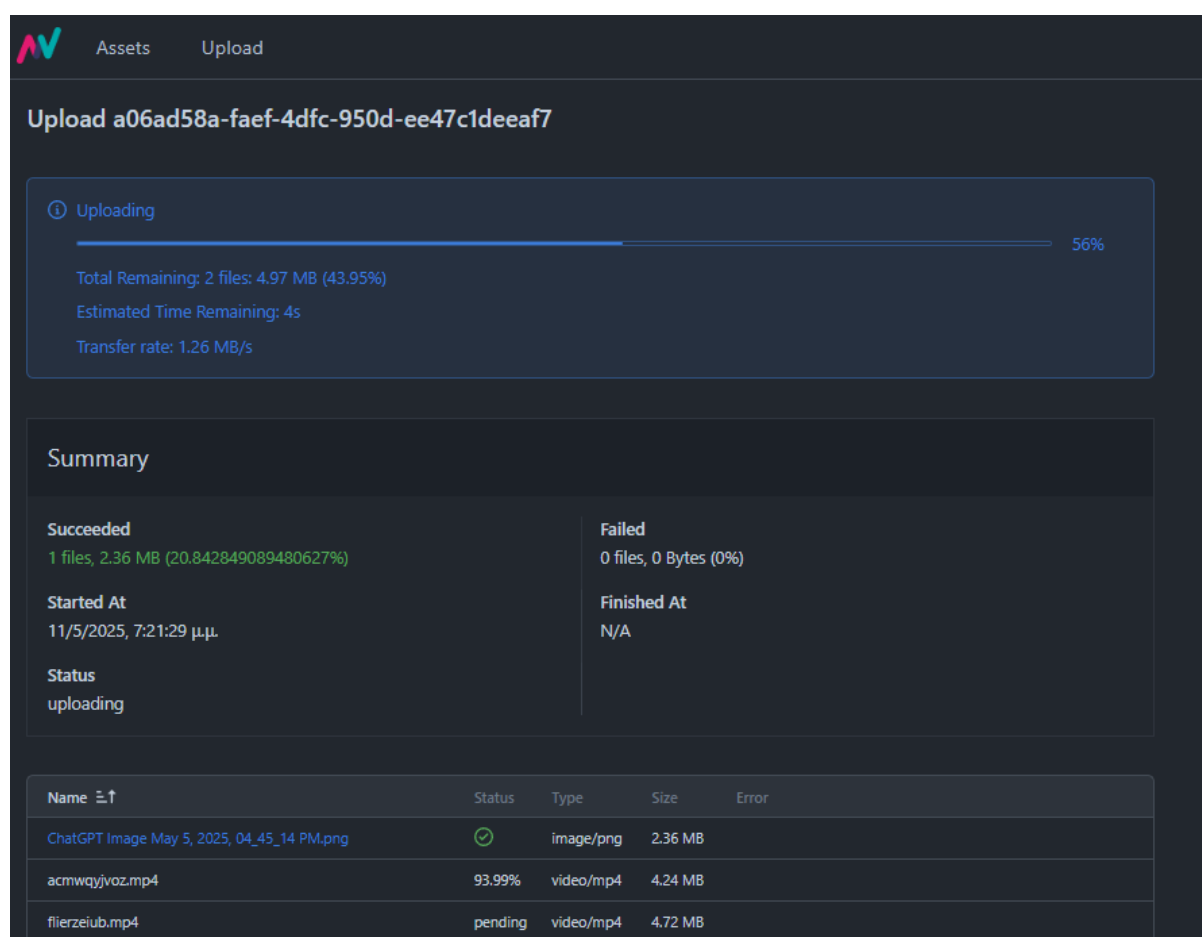
<https://maam.mever.gr/app/workspace/116/join/c6d61bbf-2b3c-4076-b601-0857d7938bb9> Copy

Users Invite

Username	Email	Name	Status	Actions
alexandrf	alexandrf@maam.mever.gr	Alexandros Papadopoulos	Active	Remove
angepara	angepara@maam.mever.gr	Angelika Papadopoulos	Active	Remove
annaniko	annaniko@maam.mever.gr	Anna Papadopoulos	Active	Remove
apsilopo	apsilopo@maam.mever.gr	Apostolos Papadopoulos	Active	Remove
artemant	artemant@maam.mever.gr	Artemis Papadopoulos	Invite Pending	Remove
chriskarak	chriskarak@maam.mever.gr	Christina Papadopoulos	Active	Remove

Figure 6 Workspace details, workspace invitation link and users participating in the workspace.

Users can upload multiple assets to MAAM by selecting files directly from their local filesystem, with the platform providing real-time progress updates during the upload process (Figure 7). In addition, MAAM supports batch uploads via CSV files, where each row contains media URLs that the system can automatically import. To further address user needs identified in the WP3 user requirements analysis, MAAM has been updated to allow users to provide links from popular social media platforms such as X (formerly Twitter), TikTok, and others. The platform automatically retrieves and imports the corresponding video content from these links, facilitating the acquisition of relevant assets for analysis.



Upload a06ad58a-faef-4dfc-950d-ee47c1deef7

Uploading 56%

Total Remaining: 2 files: 4.97 MB (43.95%)
 Estimated Time Remaining: 4s
 Transfer rate: 1.26 MB/s

Summary

Succeeded
 1 files, 2.36 MB (20.842849089480627%)

Failed
 0 files, 0 Bytes (0%)

Started At
 11/5/2025, 7:21:29 μμ.

Finished At
 N/A

Status
 uploading

Name	Status	Type	Size	Error
ChatGPT Image May 5, 2025, 04_45_14 PM.png	✓	image/png	2.36 MB	
acrmwqjyvoz.mp4	93.99%	video/mp4	4.24 MB	
flierzeiub.mp4	pending	video/mp4	4.72 MB	

Figure 7 Asset Upload Page. Users can select multiple media files for uploading.

Upon upload, media assets are automatically annotated with descriptive tags and captions that identify objects, actions, and content categories. This process also includes the detection of specialized content types, such as memes or inappropriate (NSFW – Not Safe For Work) material, allowing users to quickly understand the nature of the media they are analyzing. These annotations can be used as filters (Figure 8) during asset searches, alongside general filters such as upload date and media type, significantly enhancing content discovery. In addition to annotation-based filtering, MAAM supports free-text search across both user-provided metadata (e.g., title, description, filename) and automatically generated metadata such as AI generated captions. This combined approach ensures a highly efficient search experience, enabling users to quickly locate relevant media using keywords or specific content characteristics.

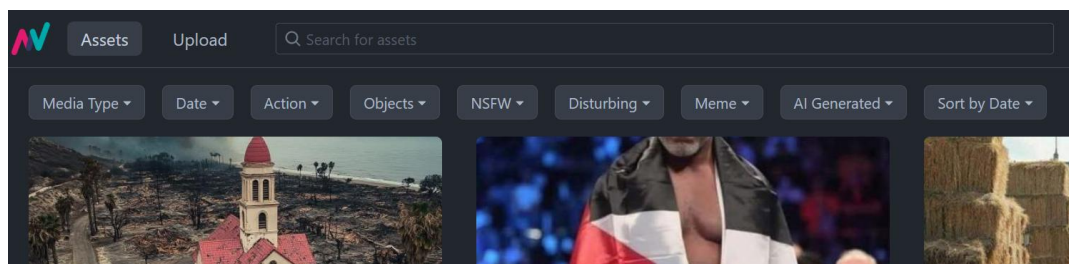


Figure 8 Filtering and Free-text search

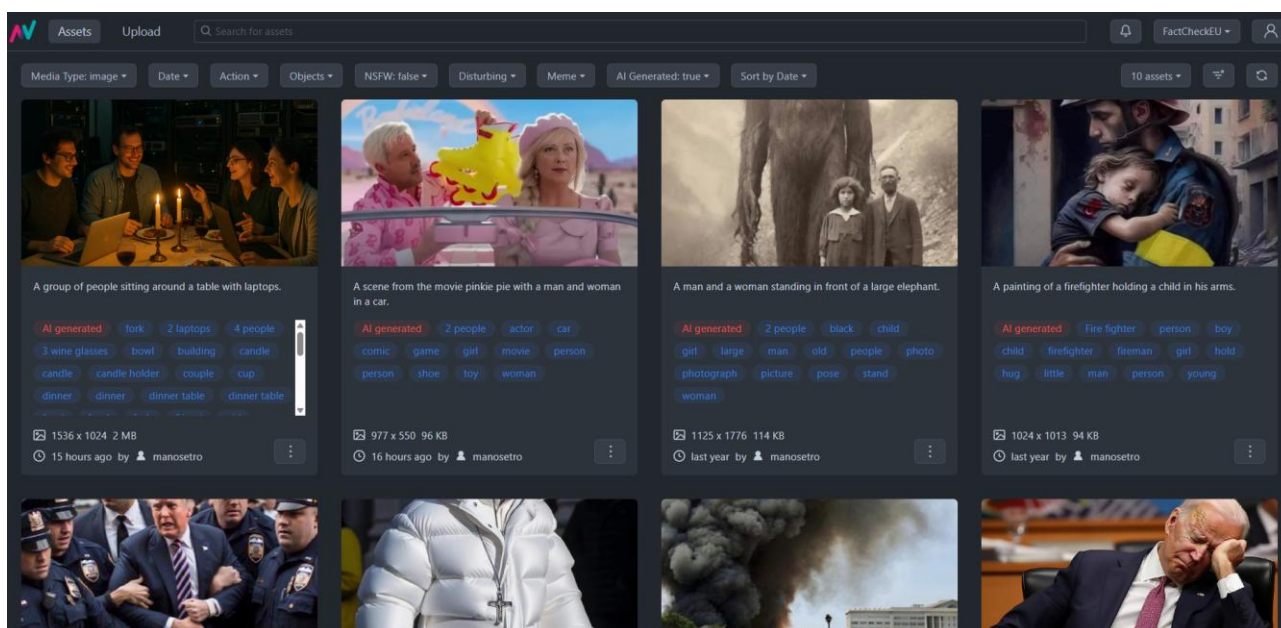


Figure 9 Assets Grid View. Assets filtered by *media_type = image*, *nsfw = false*, and *ai_generated = true*.

Figure 9 presents an overview of the assets contained within a specific workspace. Assets are displayed in a grid layout, with each asset tile showing key information such as a caption describing the asset, annotation tags, media type, and other relevant metadata. This asset view also includes a variety of filters that allow users to refine the displayed content. Users can search for specific assets by selecting predefined filters or by using the free-text search functionality, enabling quick and efficient access to relevant media. In this specific example, the user has selected images (*media_type=image*), excluding NSFW content (*nsfw=false*) including only AI generated content (*ai_generated=true*).

4.2.1 Asset Annotations with AI Models

MAAM provides a suite of image and video annotation models that are automatically applied to all uploaded assets by default. Each asset is sent to the Media Annotation Service via gRPC, and the service returns AI-generated annotations, which are then stored and indexed in Postgres and Elasticsearch respectively, as tags associated with the asset. Table 1 provides an overview of the gRPC integrated models, alongside architecture details, model size in terms of parameters and execution time. These tags enrich the metadata and enhance search and filtering capabilities. Figure 10 presents an example of automatically generated annotations within the asset details view. The annotations are organized into groups based on their type (e.g. caption, persons, etc.) allowing users to easily explore the different outputs produced by

the annotation models. Each annotation is interactive; clicking on it reveals additional details, including the confidence score associated with the prediction. In the case of images, if an annotation is linked to a specific region of the asset, this region is visually highlighted using a bounding box overlaid on the image preview on the right. In the case of videos, the temporal dimension is also taken into account for models that support tracking. As illustrated in Figure 11, annotations such as detected objects are tracked across frames, allowing the system to indicate when a specific annotation appears and when it disappears throughout the video.

Table 1 Integrated models, along with model architecture details and execution time. Execution time for images corresponds to the average processing time for 10 random images, sized between 150KB and 500KB with a resolution of approximately 1024×768.

Task		Model	Backbone Architecture	Parameters (Millions)	Execution Time (seconds)
Image	Action Recognition	ResNet152		470M	0.424
	Meme Detection	MemeTector	ViTa	3.4M	0.315
	Disturbing Content Detection	EfficientNet-B1		7.8M	0.282
	NSFW Detection	EfficientNet-B1		7.8M	0.294
	Object Detection	YOLOv8x	CSPNet	68.2M	0.313
	Face Recognition	VGGFace2	ResNet-50	25.6M	2.406
	News Tagging	Tag2Text	Swin _{Base}	109M	0.269
	Tagging	RAM	Swin _{Large}	200M	3.451
	Captioning	OFA _{Large}	ResNet-152	470M	6.645
Video	Action Recognition	SlowFast R50		34.57M	3.595
	Disturbing Content Detection	EfficientNet-B1		7.8M	4.589
	NSFW Detection	EfficientNet-B1		7.8M	4.503
	Object Detection	YOLOv8x	CSPNet	68.2M	3.159
	VideoTagging	RAM	Swin _{Large}	200M	10.688
	VideoNewsTagging	Tag2Text	Swin _{Base}	109M	1.825

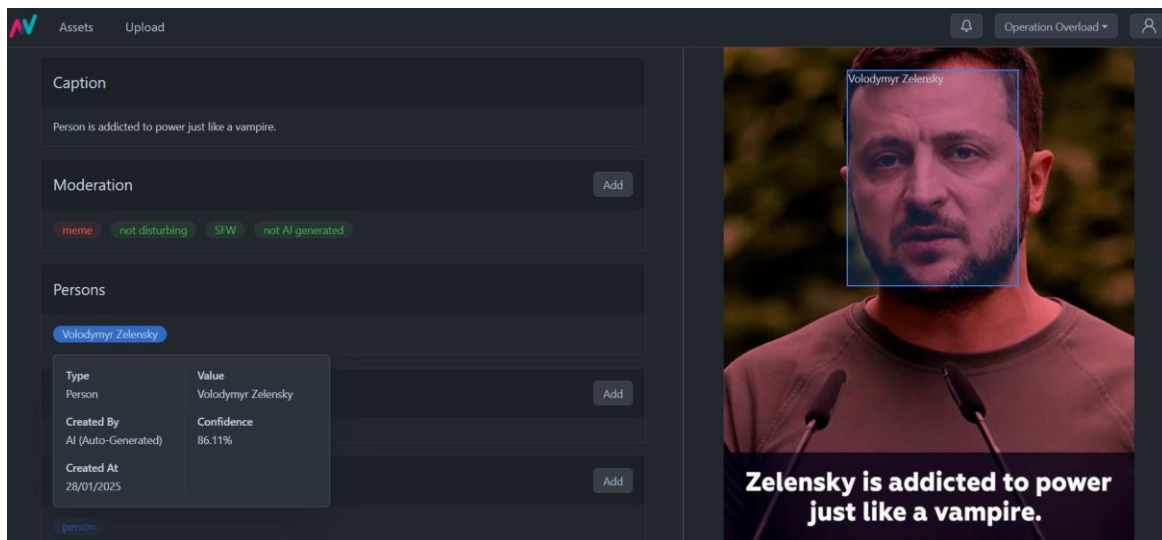


Figure 10 Asset Annotation Details for images

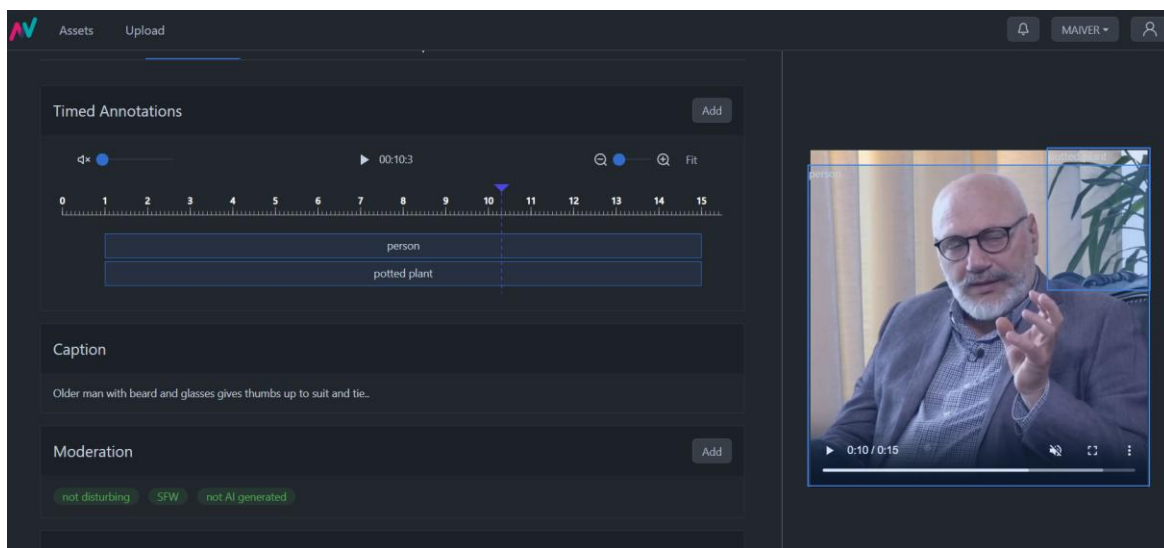


Figure 11 Asset Annotation Details for images

The annotation service includes the following annotation endpoints / models:

Automatic asset captioning: To enhance content discoverability, MAAM uses OFA [1], a state-of-the-art image captioning model, to generate descriptive text for each image. OFA unifies a wide range of cross-modal and unimodal tasks, including image generation, visual grounding, image captioning, image classification, and language modeling, within a single sequence-to-sequence learning framework. It uses a Transformer backbone in an encoder-decoder approach as a unified architecture for pretraining and fine-tuning across all tasks. In our case, we used the **OFA_{Large}** variant to balance between high accuracy and efficiency in terms of computational resources. The generated captions are indexed in Elasticsearch, making it easier to locate relevant visual content. An example of a generated caption is shown in Figure 13. For videos, OFA is applied to selected keyframes, and the resulting captions are combined into a single coherent description using the T5 language model [18]. This approach ensures that video content is also searchable, and users can have a first idea of videos content. However, this model is expected to change, as described in Section 4.3, due to its limited added value. The generated captions are often generic, prone to errors, and syntactically ill-formed.

Object detection: For object detection, the media annotation service has been updated to use a model from the Ultralytics YOLOv8 family²³. YOLOv8 offers a range of models optimized for tasks such as object detection, segmentation, and object orientation. In our case, we used the **YOLOv8x** model for the detection task. This is the largest model in the YOLOv8 family and delivers the best performance in terms of mean average precision (mAP). It is trained on the COCO dataset and can recognize 80 common objects, including people, vehicles, animals, and everyday items such as laptops and backpacks. When an image or video frame is processed, YOLOv8x identifies all visible objects and assigns each one a class label, a confidence score, and a bounding box that precisely marks its location. The same approach is applied on videos at selected keyframes.

Action recognition: For action recognition in videos, we use the SlowFast R50 model [2] trained on the Kinetics400 dataset [3]. For images, since Kinetics400 is a video dataset and most action recognition models rely on 3D CNNs to consider the temporal dimension, we adopt a ResNet152 backbone [4] with the TSN approach [5], for training and inference at frame level.

Media tagging: This annotation endpoint provides two complementary sets of tags that describe the content of each image. The first is based on the Recognize Anything Model (RAM) [6], a large-scale image tagging model capable of identifying up to 6,449 predefined semantic tags, offering a broad and detailed description of the image content. The second model is derived from the Tag2Text model [7] and has been fine-tuned on public news-related image datasets. This tagging model generates news-oriented tags, providing more contextually relevant annotations for media assets typically encountered in journalism and reporting.

Face Detection and Recognition: The face recognition method combines MTCNN²⁴ for face detection and the InceptionResNetV1 architecture from FaceNet [8] for generating face embeddings. The face encoder is trained on the VGGFace2 dataset [9]. To build a reference database, we used the DBpedia Lookup API²⁵, to select approximately 52,000 well-known individuals Wikipedia entries based on page view statistics. For each selected identity, we collected high-quality images, extracted faces, and computed facial embeddings. The final database consists of over 328,000 images representing approximately 46,000 unique identities. This streamlined approach can also be used to periodically extend the reference database with new individuals as they gain relevance. Given an uploaded asset the system detects all the faces and encodes them using the same face encoder. Each query embedding is then compared against the reference database using cosine similarity and the most similar identities whose similarity exceeds a predefined threshold, are returned. For videos, the same approach is applied to sampled keyframes.

Content Moderation: MAAM uses two moderation models to detect NSFW and disturbing content, trained via a semi-automated annotation process on large image datasets [10]. These models are based on the EfficientNet-B1 architecture, as it has been shown to be one of the most effective convolutional neural network (CNN) architectures across a variety of visual

²³ <https://docs.ultralytics.com/models/yolov8/>

²⁴ <https://github.com/timesler/facenet-pytorch>

²⁵ <https://www.dbpedia.org/resources/lookup/>

tasks. For images, moderation scores and tags are added at image level. For videos, keyframes are analyzed, and a video is flagged as NSFW or disturbing, if any scene is tagged. Users can filter such content during search, and the UI blurs flagged assets by default using artistic style transfer, allowing users to reveal them if desired. Figure 13 shows an example of an asset detected as disturbing.

Meme Detection: To identify memes, we use our custom-developed MemeTector [11] model, that provides also a confidence score. MemeTector has been trained based on a methodology that treats the visual components of image memes as instances of the *regular image* class, while the original image memes are treated as instances of the *meme* class. This approach encourages the model to focus on the critical visual elements that distinguish a meme from a regular image. It employs a trainable attention mechanism on top of a standard Vision Transformer (ViT) architecture, enhancing the model's ability to focus on these features and making predictions more interpretable.

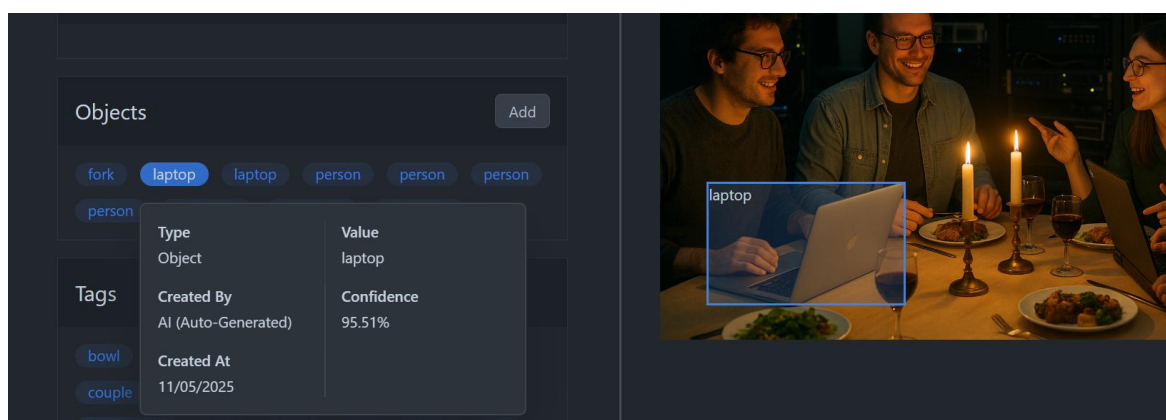


Figure 12 Object detection and localization

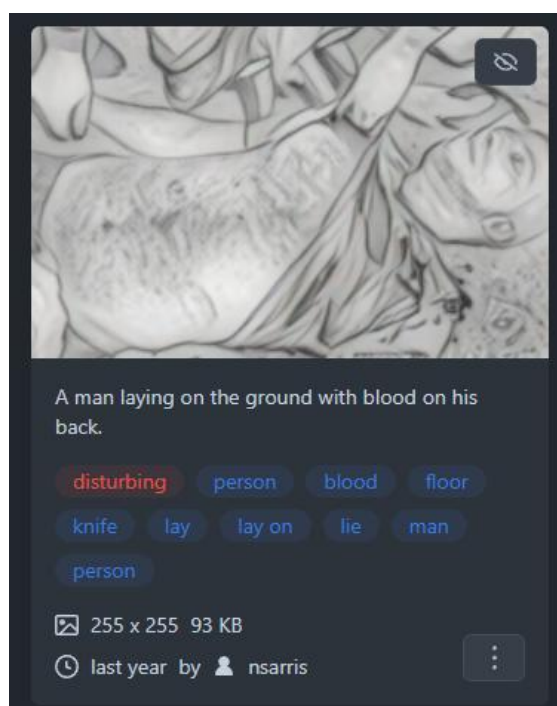


Figure 13 Image tagged as disturbing, with style transfer hiding the content to prevent exposure of users

to it. Captioning provides also a textual description of the content.

The location estimation tool is a service integrated into MAAM, which, given an image, provides an estimate of its location based on the visual analysis of its content e.g., based on landmarks or visual cues that are specific to a specific region. The method which is based on and extends the work of [27], consists of two steps and leverages a reference dataset of over 4 million geotagged images. First, a classification model predicts which spatial cell an input image belongs to, using Google's S2 Geometry library²⁶ to divide the Earth's surface into rectangular regions based on geographic coordinates. Then, in the retrieval step, the system searches within the predicted cell for the most visually similar image using cosine similarity. The estimated location information is available in the asset details. The tool returns the most likely location for the image, shows the confidence level of the estimation, identifies the location name, shows a pin with the location in a map, and displays some photos from the identified location as illustrated in Figure 14.

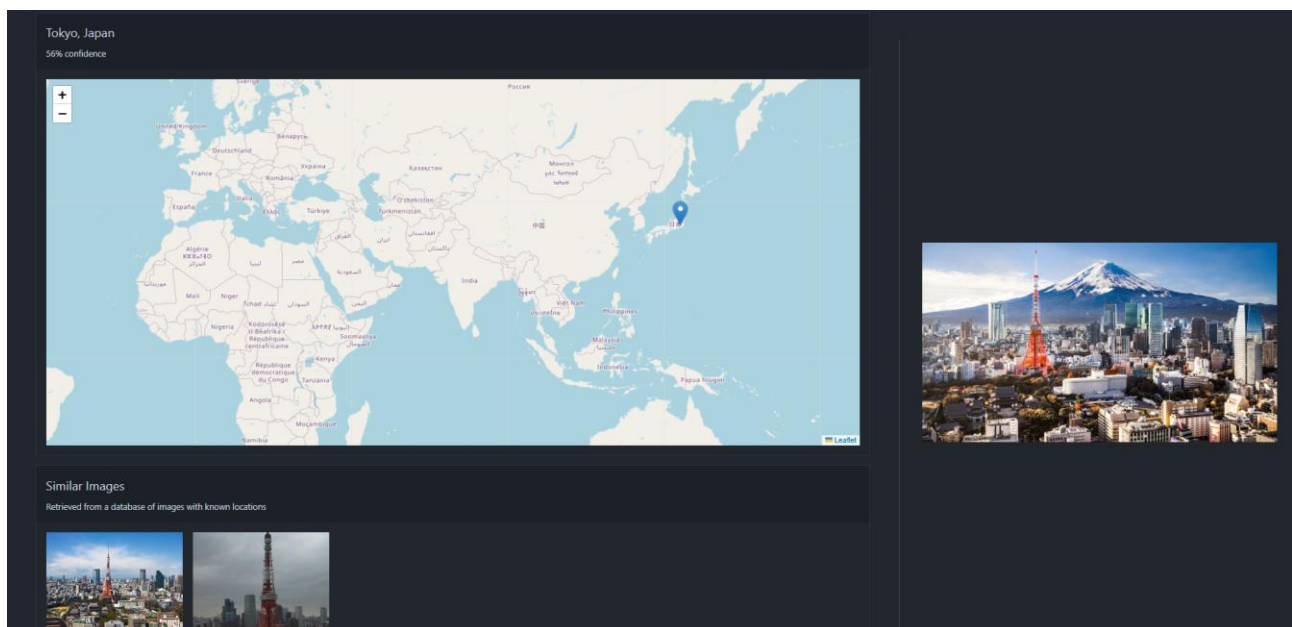


Figure 14 Geolocation estimation on images

4.2.2 Visual similarity and near duplicate detection

To enable image-based visual similarity search, MAAM leverages a pre-trained CLIP model [12], trained on approximately 400 million image-text pairs. Specifically, we use the ViT-B/32²⁷ variant of the image encoder. For each image uploaded to the DAM, MAAM obtains a 512-dimensional dense vector from the Media Annotation Service that captures the semantic content of the image. This vector is then indexed in Elasticsearch to support similarity-based retrieval using its approximate k-nearest neighbor (kNN) search capabilities. Elasticsearch provides a field type for dense vectors²⁸ optimized for efficient retrieval of similar vectors through kNN searches. This kNN search can be seamlessly combined with the other Elasticsearch features, such as free-text search, filters, and aggregations, allowing users to

²⁶ <http://s2geometry.io/>

²⁷ <https://huggingface.co/openai/clip-vit-base-patch32>

²⁸ <https://www.elastic.co/docs/reference/elasticsearch/mapping-reference/dense-vector>

refine and tailor their similarity queries for more relevant results. Beyond global image similarity, MAAM also supports region-based retrieval. Users can select a specific area within an image using a bounding box, and MAAM generates a dense vector representation of that region in real time. This region-level vector is then used in the same kNN search process to retrieve visually similar content, enabling more precise and focused discovery. Figure 15 provides an overview of the workflow for the extraction and indexing of CLIP embeddings. Figure 16 provides an example of visual similarity. Note that the retrieved images are not only visually similar but also semantically related, reflecting the ability of CLIP to capture high-level concepts.

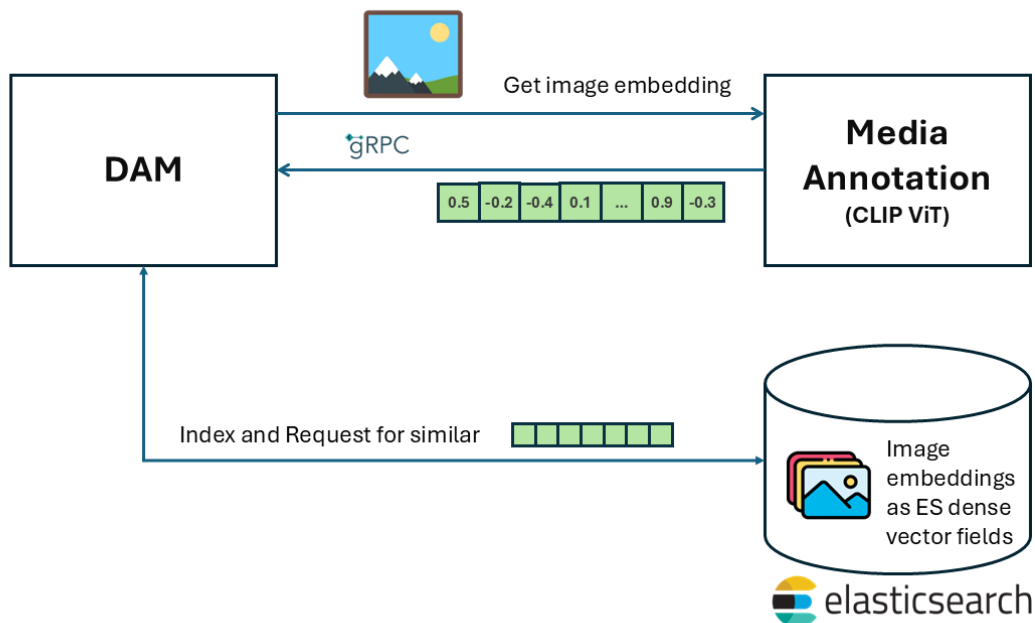


Figure 15 Workflow for Indexing CLIP Vectors to Enable Visual Similarity Search

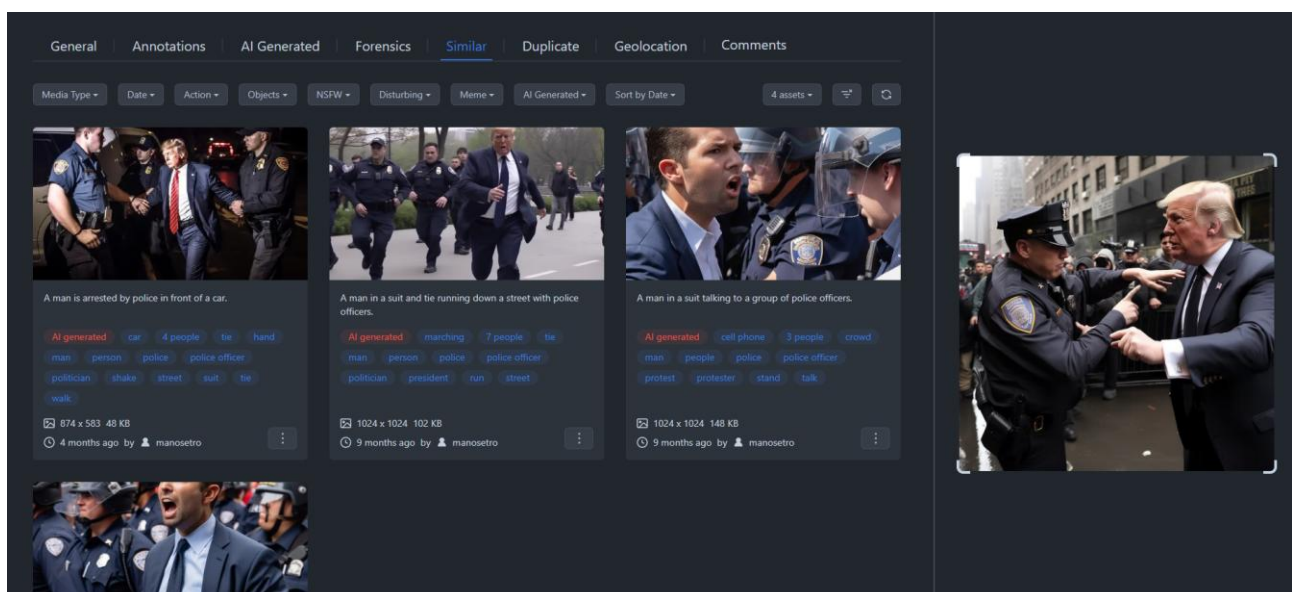


Figure 16 Image visual similarity using CLIP embeddings to capture semantic information. Image on the right is used as a query, while the images on the left are the visually similar results.

In addition to visual similarity, the Near-Duplicate Detection (NDD) service offers an efficient reverse search functionality for both images and videos, allowing users to quickly identify duplicate or near duplicate assets within their collections. Built on the Distill-and-Select (DnS) method [13], the NDD service ensures high accuracy and scalability in both indexing and retrieval tasks. DnS is a knowledge distillation approach that starts from a strong fine-grained Teacher Network and learns multiple Student Networks that offer different trade-offs between retrieval performance and computational efficiency. Alongside the students, a Selector Network is also trained to quickly choose the most appropriate student for each input during inference, ensuring both high performance and low computational cost. The student models vary in architecture and efficiency, including lightweight versions that use binary representations, enabling faster retrieval and reduced storage requirements.

While it supports indexing and querying duplicate images, such as those that have undergone filtering, cropping, or other modifications, its main strength lies in handling videos. The key advantage of the service is its ability to detect video-level duplications, including cases where a segment of one video is reused in another, or where a video is embedded as a background or partial overlay in a different media asset. When a media item is submitted, the system analyzes its visual content and adds it to a dedicated index. Users can then perform similarity-based queries to retrieve and rank near-duplicate results, making the service particularly valuable for uncovering reused or repurposed video content across large datasets.

NDD is a standalone service and supports several key operations:

- i. Adding images and videos to the index via explicit media URLs
- ii. Querying the index for near-duplicates with configurable similarity metrics, including video-level and short level search
- iii. Creating and managing collections of media items for fine-grained search

Internally, the NDD service follows a modular, service-oriented architecture, with separate components for feature extraction, indexing, and searching. It communicates with the MAAM platform through a dedicated REST API, ensuring seamless integration and interoperability.

In the current context of AI-CODE, we are working to integrate segment-level comparison functionalities to identify duplicate or near-duplicate segments between videos, which may result from modifications made for disinformation purposes.

As NDD supports the concept of collections, to align with MAAM's organizational structure we map directly NDD's collections to MAAM workspaces. For each workspace created in MAAM, a corresponding collection is created in the NDD service. When assets associated with a workspace are added to NDD, they are indexed within that specific collection. This design ensures that duplicate search operations are scoped to the workspace, preventing cross-workspace retrieval.

4.2.3 Media Asset Verification Tools

The DeepFake Synthetic Image and Video Detection Service in MAAM offers a list of detection models trained to identify images generated by state-of-the-art generative algorithms, such as GANs and diffusion models. Each detector is optimized to recognize artifacts specific to

specific generation techniques, meaning they are most effective at identifying content produced by certain types of generative models. By combining multiple detectors, the system provides users with a stronger and more reliable signal to assess the likelihood that an image is AI-generated. These detectors are calibrated to maintain a low false positive rate, so when an image is flagged as synthetic, there should be a very low probability of the image being real. Based on this calibration, the detectors assign an *evidence level* to each prediction (Table 2). This, combined with a clear explanatory message, presents the detection result in a human-friendly way that is easy for non-experts to understand. For example, a label such as “**very strong evidence**” indicates that the likelihood of a false detection is below 1%, providing users with clear, guidance in relation to the reliability of the specific result. Figure 17 presents the results produced by the detection models when applied to an AI-generated image of Pope Francis wearing a pride flag²⁹. These images, which had also been shared in previous years, began circulating more widely after Pope Francis’s death, particularly on X (formerly Twitter), and were often used to portray him in a negative light. Many of these posts were accompanied by hate-filled captions, with the late pontiff being labeled as “irresponsible” and “perverted”³⁰. The figure from MAAM results illustrates how the detectors assess the likelihood of the image being synthetic, providing a clear example of their output in a real-world scenario.

Table 2 Evidence levels in AI generated content detection

Evidence	Message
Very strong evidence	Very strong evidence suggests AI generation, with a false positive rate below 1% .
Strong evidence	Strong evidence suggests AI generation, with a false positive rate from 1% to 5% .
Moderate evidence	Moderate evidence of AI generation detected, with a false positive rate up to 10% .
Weak evidence	Weak evidence of AI generation detected, with a false positive rate over 10% .

²⁹ <https://cedmohub.eu/images-of-pope-francis-wearing-pride-flag-are-ai-generated/>

³⁰ <https://meddmo.eu/ai-images-of-pope-francis-wearing-pride-flag-resurface-after-his-death/>

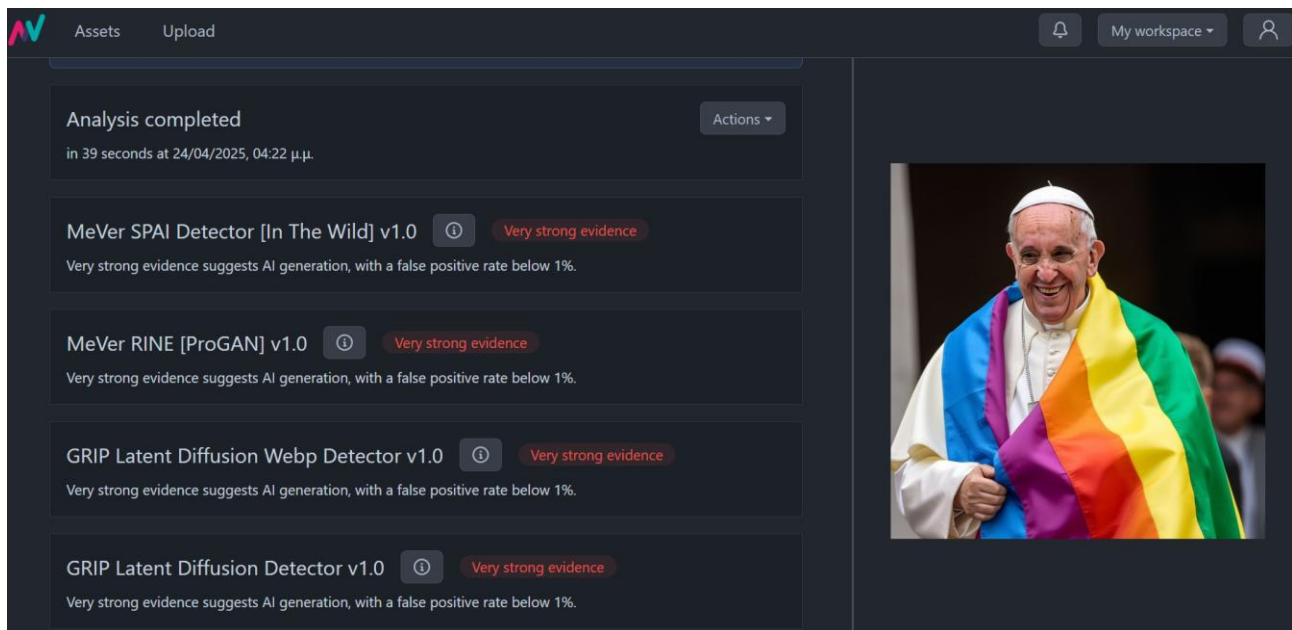


Figure 17 AI Generated Analysis for images. This tab provides the results of the Deepfake and Synthetic Images detection service. In the case of images, the results are a list of predictions from the detection models integrated into the service.

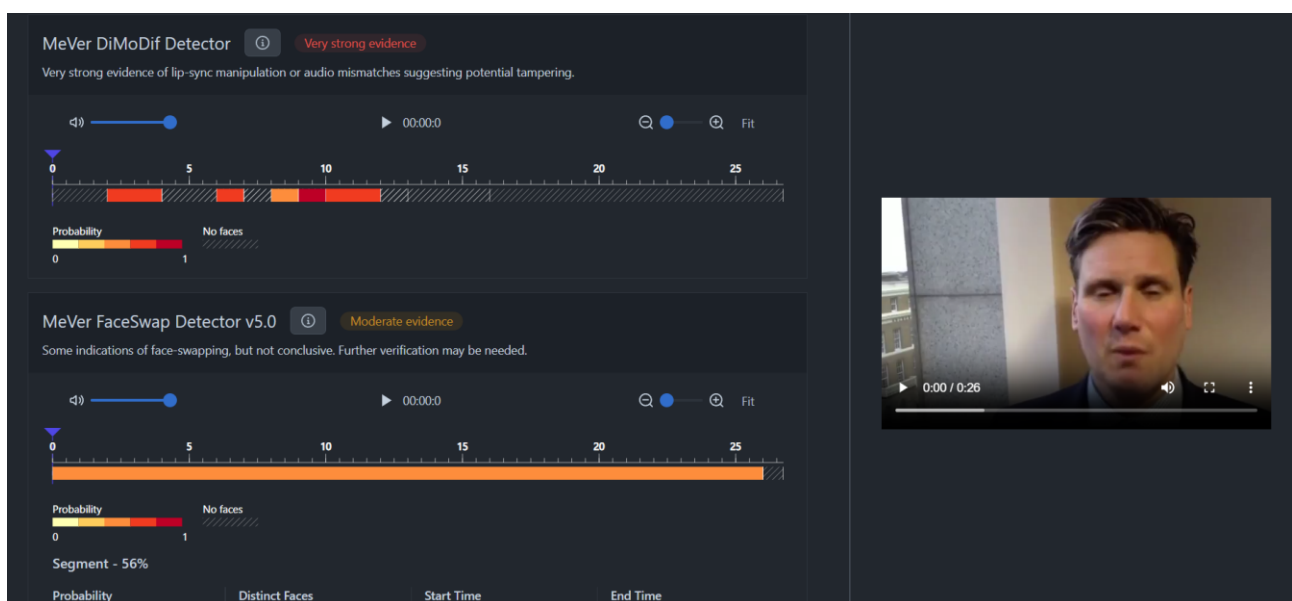


Figure 18 AI Generated Analysis for videos.

For video analysis, MAAM currently incorporates two DeepFake detection models. The first is trained on FaceSwap generated content and is optimized for detecting identity replacement, a common technique in DeepFake videos. The second, DiMoDif [28], is specifically designed to detect lip-sync manipulations by identifying mismatches between a speaker's lip movements and the corresponding audio. While DiMoDif is not intended for general synthetic video detection, it effectively targets audio-visual inconsistencies in cases where the speaker is clearly visible and audible. Together, these models provide complementary capabilities, enhancing MAAM's effectiveness in detecting a broad range of DeepFake video manipulations. In the same way as synthetic image detectors, DeepFake detectors provide evidence level, and corresponding messages to help users understand the likelihood of a video being manipulated.

Figure 18 presents an example of a deepfake video featuring British Prime Minister Keir Starmer. The manipulated video falsely suggests that he was aware of the investigation into paedophile Jimmy Savile but chose to halt it, dismissing the claims as “frivolous.” The deepfake retains the original visual content but replaces the audio to alter the meaning of the statement, creating a misleading narrative. The MeVer DiMoDif Detector results suggest the video has most probably been manipulated through face reenactment techniques.

The Forensic Analysis tab in MAAM offers tools for detecting image manipulations such as splicing, where elements from different images are combined, and copy-move forgeries, which involve duplicating and relocating parts within the same image. It also supports detecting manipulations introduced by text-guided inpainting methods (such as Stable Diffusion 2 [14], Stable Diffusion XL [15], and Adobe Firefly³¹), where modified regions are spliced into the original image. These manipulations are visualized through forensic heatmaps, which highlight suspicious areas to help users assess image authenticity. The forensic algorithms have been extended to include state-of-the-art detectors, such as TruFor [17], improving the tool’s ability to localize manipulated regions. However, the tool has still limited effectiveness in identifying tampering in fully regenerated images, where the entire image has been recreated. An example from the TGIF dataset [16], which includes such inpainting-based manipulations, is shown in Figure 19. In this example, spliced, in-painted areas, can effectively be detected, by state-of-the-art forgery localization models such as TruFor [17].

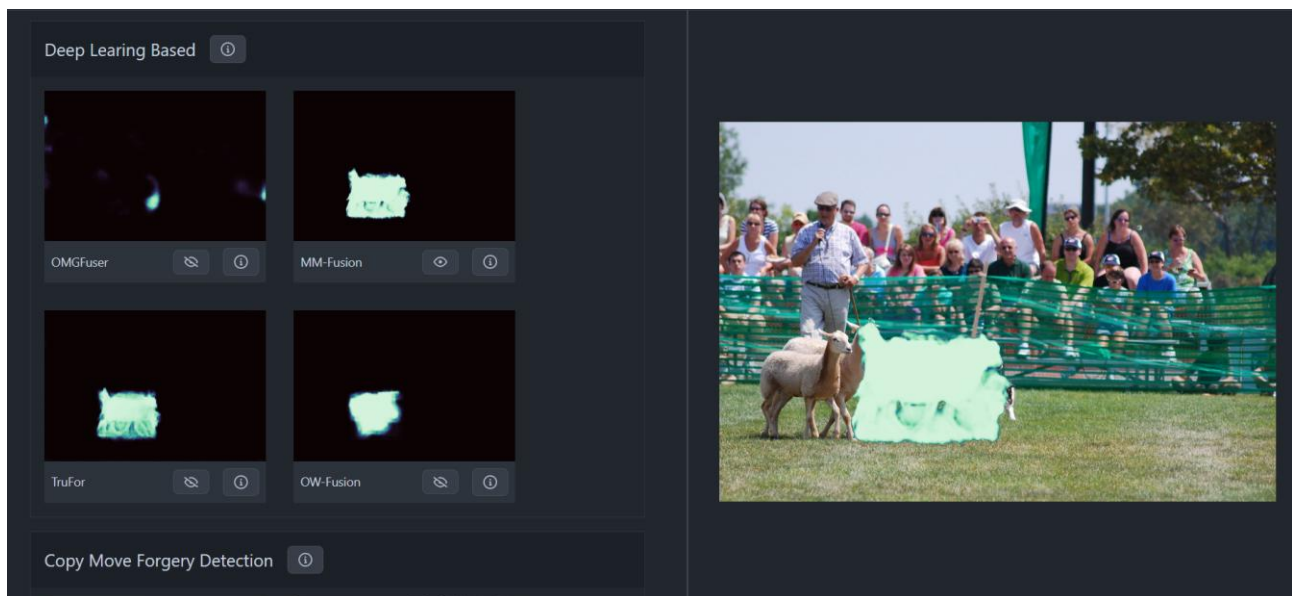


Figure 19 Forensics Analysis. The forensic heatmaps generated by the integrated methods (shown on the left) are overlaid on the image (shown on the right) to help users identify regions that are most likely to have been forged.

4.3 Model Soundness

For most annotation tasks, and particularly in the case of images, the integrated models deliver state-of-the-art performance and are considered satisfactory at this stage. However, the main

³¹ https://www.adobe.com/gr_en/products/firefly.html

limitations have been identified in relation to video captioning. Due to the current frame-based analysis approach, which overlooks the temporal dimension, the effectiveness of the integrated models for video annotation is reduced. As a result, video captioning stands out as the primary video annotation task that requires improvement, since understanding the temporal sequence of events is crucial for generating meaningful and accurate descriptions.

In the context of previous projects that laid the groundwork for MAAM development, the need to automatically generate accurate and meaningful video captions led us to explore state-of-the-art and robust solutions in multimodal learning and video captioning. We initially experimented with the OFA [1] model by extracting one frame per second from videos, captioning each frame independently, and then summarizing the sequence using a T5 model [18]. This is the model currently integrated into MAAM. However, this approach fails to capture the temporal flow of video content and performs poorly especially in long videos. Although a stronger video model, Video-OFA demonstrates promising performance [19], it is not open-source and thus not usable in our setting. This limitation led us to explore Vision-Language Models (VLMs) better suited for video understanding.

After evaluating the available options in the related literature, we adopted LLaVA-OneVision [20], a state-of-the-art open-source VLM that offers strong performance, flexibility, and efficiency in vision-language tasks, with strength in bridging the gap between video understanding and natural language generation. LLaVA-OneVision belongs to the LLaVA family [21] of large multimodal models (LMMs), designed as a general-purpose visual assistant. It can be used in tasks such as video question answering and video annotation, but it is particularly well-suited for video captioning thanks to its simple but effective architecture and training methodology, which make it capable of handling both static and dynamic visual content.

Its architecture combines a vision encoder, particularly SigLip [22], to extract visual features from the input frames, with a large language model (LLM), Qwen [23], known for its strong language capabilities. This connection is facilitated by a lightweight projection module (a 2-layer MLP projector) which connects the visual features to the LLM's word embedding space, making multimodal understanding easier.

The model was trained in three main steps: first, it learned to connect visual and language features; then, it improved its reasoning using high-quality image-text pairs, and finally, it was fine-tuned on video-caption and instruction-based data to better understand video content. The fine-tuning involves two key phases: (i) on 3.2 million carefully selected single-image samples to build strong multimodal capabilities, and (ii) on an additional 1.6 million mixed samples (videos, multi-image sets, high-quality single images), to help the model handle both static and dynamic visual content.

LLaVA-OneVision demonstrates strong performance across various video captioning benchmarks such as ActivityNet-QA [24], VideoDetailCaption [20], PerceptionTest [25], VideoMME [26] and others (Table 3). Its captions effectively capture both visual details and the temporal progression of actions and scenes, essential for producing accurate and meaningful captions. The model's architecture enables task transfer, allowing it to generalize from single image to complex video captioning tasks efficiently, even with limited computational resources.

Its results are competitive with, and sometimes surpass, advanced models like GPT-4V³², which are not open-source.

Table 3 LLaVA-OneVision performance on video benchmarks. For VideoDC results are presented as ratings out of 5. All other benchmarks are reported in terms of 0-shot accuracy (%).

Benchmarks	LLaVA-OV-7B	LLaVA-OV-72B	GPT-4V
ActivityNet-QA	56.6	62.3	57
VideoDetailCaption	56.7	59.4	43.5
PerceptionTest	57.1	66.9	-
VideoMME	58.2 / 61.5	66.2 / 69.5	59.9 / 63.3
VideoDC (VideoDetailCaption)	56.7	59.4	43.5
MVBench	64.7	68.0	49.2

Although LLaVA-OV-72B achieves the highest performance across most benchmarks, in our work we employ the LLaVA-OV-7B variant due to computational constraints, as it offers a better trade-off between efficiency and accuracy for video captioning tasks. In addition to its strong benchmark results, LLaVA-OneVision is practical and efficient. Its modular design and lightweight projection head allow for fast inference, making it a valuable tool for real-time or large-scale subtitle generation and video understanding tasks in resource-constrained environments. Finally, LLaVA-OneVision provides a solid and practical foundation for developing video captioning systems that meet real-world demands in terms of accuracy, scalability, and efficiency. Its multimodal architecture, combined with fine-tuning on video-caption and task-specific data, makes it a highly effective tool for generating captions in diverse types of video content.

4.3.1 Video captioning with LLaVA-OneVision

To use LLaVA-OneVision for captioning, we provide a generic prompt alongside the input video. Additionally, we include any detected persons as auxiliary signals to guide the captioning process. Figure 20 depicts selected video frames from a video used as an example. The OFA-based approach with summarization using T5 provides a generic caption:

“Woman delivers speech in front of a flagpole. flagpole is decorated with red, white and blue flowers”,

which is inaccurate and syntactically problematic. In addition, it does not provide any additional information compared to the tag annotations. On the other hand, the prompt-based LLaVA-OneVision uses this:

“Provide a clear and engaging description of the video showing [Ursula Von Der Leyen]. Describe only what is visually present. No assumptions, sounds, or unseen events.”

to get the following response:

“The video features Ursula Von Der Leyen standing at a podium with a microphone, speaking to an audience. She is dressed in a professional yellow attire, with a backdrop

³² <https://openai.com/index/gpt-4v-system-card/>

displaying the European Union flag and the European Union logo. The setting suggests a formal event, possibly a press conference or a public address.”



Figure 20 Video example for captioning

It is important to note, however, that **LLaVA-OneVision** is highly sensitive to the input **prompt**, which can lead to hallucinations. For example, using the prompt:

“Provide a detailed, clear, and engaging caption for the video. Make sure to include the most important details and avoid unnecessary information. The person in the video is [Ursula Von de Leyen]”

can cause the model to generate content that goes beyond visual evidence, resulting in hallucinations. A typical output for this prompt is:

“Ursula von der Leyen, the President of the European Commission, addresses the European Parliament on the state of the union. She outlines the key priorities for the commission, including strengthening the European Union’s role in global affairs, promoting democracy and human rights, and ensuring the protection of the environment.”

This caption includes information about the speech that is not visually verifiable from the video itself and is clearly hallucination introduced by the language model, based on the contextual association with the name of Ursula Von de Leyen. Therefore, before integrating such a model in MAAM, additional evaluation of video annotation is needed, including user-based assessment to be conducted as part of WP8’s validation methodology.

4.4 Ethical Design Considerations

Ethical considerations have played a central role in the design and development of MAAM, particularly due to its focus on analyzing and verifying media content that may involve human subjects, sensitive material, and AI-generated outputs. The development of MAAM has been guided by the principles of Trustworthy AI, based on the Ethics Guidelines for Trustworthy AI by

the High-Level Expert Group on AI (AI-HLEG)³³. This work has been conducted in collaboration with Work Package 9, particularly with Task 9.3, which focuses on the development and implementation of an ethical assessment framework for all AI-CODE services. Some of these ethical requirements were already addressed in the user requirements analysis of Deliverable D3.1, and the preliminary analysis has since been revisited and refined considering the ongoing system development.

It is also important to note that many of the ethical concerns related to Trustworthy AI, such as bias, fairness, and transparency apply not only to MAAM itself but also to the underlying AI models integrated into MAAM. As MAAM serves primarily as an integration point for various verification models developed across WP4 and WP5, ethical considerations extend to those components as well. Table 4 highlights the main ethical and human oversight requirements identified through this process and explains how MAAM addresses them through specific design and implementation choices. The requirements are categorized based on the seven requirement categories of Trustworthy Artificial Intelligence (AI), as specified in Ethics Guidelines for Trustworthy AI:

1. Human Agency and Oversight (HAO)
2. Technical Robustness and Safety (TSA)
3. Privacy and Data Governance (PDG)
4. Transparency (TRA)
5. Diversity, Non-discrimination and Fairness (FAI)
6. Societal and Environmental Well-being
7. Accountability

Table 4 Ethical design considerations based on the requirements for Trustworthy AI. These considerations cover the 5 of the 7 categories

# Req.	Ethical requirements	Design and Implementation
HAO-01	It shall be apparent to the user interacting with the application that they are interacting with an AI system.	The interface now provides detailed information about how the integrated AI models operate. For each outcome presented to the user such as annotations or detections, there is an explicit mention of the specific AI model used to generate it. This approach improves transparency and ensures that it is clear to users that AI systems are embedded in the service. This will further be improved with the use of models cards, instead of static descriptions for the AI models.
HAO-02	The capabilities of the application shall be communicated clearly to the user.	Each result, whether it is an annotation or the output of a verification tool, is accompanied by an 'info' icon.
HAO-04	The application shall take into account the risk of stimulating dependency in users.	By explaining how each AI model works, MAAM encourages users to stay critical and avoid over-reliance. The use of multiple detectors with distinct evidence levels highlights

³³ <https://op.europa.eu/en/publication-detail/-/publication/73552fcd-f7c2-11ea-991b-01aa75ed71a1>
<https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1/language-en>

		model limitations and inconsistencies, reinforcing the need for human oversight in the verification process.
TSA-02a	The accuracy, reliability, and reproducibility of the application shall be monitored, documented, and communicated to the user over time.	Steps are being undertaken to integrate model cards from T6.2 and T7.2 in MAAM.
TSA-02b	There are processes in place to ensure that the level of accuracy of the AI system to be expected by end-users and/or subjects is properly communicated	
TSA-03	The data used to build the application shall be up-to-date, of high quality, complete and representative of the deployment environment.	MAAM's modular architecture allows for easy integration of up-to-date models trained on recent high-quality datasets. This ensures the system remains relevant to real-world conditions and helps prevent performance degradation caused by data distribution shifts over time.
PDG-01	The application shall minimize the amount of data collected on users.	No personal user data is collected beyond the email address and organization name provided during registration. Currently, no user profiles are created, and no further profiling or tracking of personal information takes place within MAAM.
TRA-02	The decisions of the application shall be understandable by the user (XAI) as far as reasonably achievable.	The user can request information on e.g. underlying definitions of the integrated AI models. Model cards will be integrated at a later stage with TSA-02a and TSA-02b.
TRA-03	The application shall inform users about the purpose, criteria, and limitations of the decisions generated by the AI system.	
FAI-01	The application shall have procedures in place that avoid unfair bias that apply to data, design, and output.	Bias mitigation strategies will be applied during the training and implementation of detection models, with fairness checks integrated into the evaluation process. The transparency service will document all relevant details, including dataset composition and model performance, to ensure visibility into potential sources of bias.
FAI-03	The application shall have a mechanism that allows for flagging issues related to bias, discrimination or poor performance.	To address this ethical requirement, a user feedback mechanism is currently under development in MAAM. This feature will allow end users to rate AI-generated annotations, provide input on the detection results of synthetic image and DeepFake detectors, and comment on forensic analysis outcomes. In future iterations, this mechanism will be expanded to include feedback specifically related to issues of bias, discrimination, or poor performance, enabling ethical oversight of the system's outputs.
HAO-05	The level of human oversight (HITL, HOTL, HIC) is clearly defined and appropriate for the	The user receives a recommendation from the system. The system is not agentive, thus the oversight level remains HIC and is appropriate for the use case.

	use case	
--	----------	--

In addition to the Trustworthy AI considerations, feedback provided by the Ethical Advisory Board (EAB) of the AI-CODE project has been incorporated into the development process (Table 5). The EAB plays a key role in ensuring the ethical integrity of the project by identifying potential shortcomings in current practices and offering recommendations, best practices, and practical examples to guide improvements. EAB's input has helped refine MAAM's alignment with the broader ethical framework of the project.

Table 5 Ethical Advisory Board recommendations

EAB recommendation	Implementation measure
Users must be informed about the Terms of services.	To ensure informed consent, a pop-up window will be implemented upon launching the platform that presents the Terms of Service, tailored to the specific workspace type (private, shared, or public) the user is entering.
Clarifications needed for potential users: • Where their uploaded data is stored. • Who can access the uploaded assets. • Whether their data is used for further training.	The MAAM platform has been designed with the following principles in mind: • Data Storage: uploaded content is stored securely in protected databases. • Access Control: Each asset is associated with a workspace (private, shared, or public as selected by the user) and access is managed accordingly. • Model Training: There is currently no retraining of AI models based on user-uploaded content or user feedback. The model training policy is under consideration, as explicit user feedback may serve as a valuable input for future improvements of the detection models. If adopted, such feedback would be provided consciously by the end users, with clear consent that their ratings may be used for fine-tuning purposes. No implicit data collection from user interactions will be used. In all cases, best practices for transparency, and user privacy will be applied.
The user should have an option to permanently delete content they upload.	The MAAM platform already supports full data deletion. Users can delete uploaded content, metadata, system-generated predictions, and any feedback they have provided about the asset analysis. If the handling of user feedback is revised in future versions of the platform, it will be subject to a dedicated ethical review and consultation process.
Enhance the explicability of system outputs by creating more transparency on the performance of the models.	As part of Tasks T6.2 and T7.2, model cards will be implemented to document and communicate the limitations, intended use cases, and performance characteristics of the annotation models and integrated detectors used within MAAM.
Make the user aware of potential pitfalls and weaknesses in the system, such as levels of accuracy, to ensure an appropriate level of reliance on the system.	These model cards will be fully integrated into the MAAM platform to provide users with transparent, accessible insights into how each AI tool operates supporting informed interpretation of the results.

If user profiles or interaction data are analyzed to improve the functionality of the MAAM platform, such analysis should be conducted in a way that avoids profiling of users and usage of sensitive characteristics about them.	At this stage, there is no user profiling within the MAAM platform. Any monitoring of usage or logging activities will be strictly limited to user interactions with the platform (e.g., feature usage, workflow patterns) and does not involve the collection or analysis of personal user data. However, this approach might be reassessed as future decisions are made regarding potential improvements to the system.
---	---

5. Conclusions

This deliverable presents a prototype called MAAM (Media Asset Annotation and Management), developed to support the needs of the AI-CODE project in the domain of media verification, annotation, and content management. The prototype integrates a comprehensive set of functionalities tailored for the annotation, organization, retrieval, and verification of media assets, including both images and videos. These functionalities have been designed in alignment with the identified requirements of the project and reflect the outcomes of activities such as co-creation workshops and feedback sessions.

The current version of the MAAM prototype already supports a wide range of real-world use cases to a substantial extent. These include use cases related to the detection of synthetic media, deepfake identification, moderation of harmful content, retrieval of similar or near-duplicate assets, and enrichment of multimedia assets with semantic labels. It also incorporates a first version of tools for manual annotation, and user-driven feedback, enhancing the human-in-the-loop workflow that is essential for trustable media analysis.

6. Next Steps

While the system is already functional and usable, ongoing development efforts are focused on improving scalability, user experience, integration of more advanced detection models, and better explainability of AI outputs. These improvements aim to further increase the system's relevance and usability for a diverse set of end users, ranging from journalists and fact-checkers to social scientists, researchers, and media professionals.

The next steps in the development of MAAM focus on expanding its verification capabilities, improving user experience, and deepening the integration of ethical and explainable AI principles. This work will be carried out in WP5 (T5.3), which builds upon and continues the efforts initiated in WP4 (T4.3). One of the immediate priorities is to enhance how detection results are visualized by introducing an overall assessment layer in the detection views. This improvement responds to the requirements analysis from WP3, which highlights the need for multi-level detail views to serve both expert users and non-experts. In this context, a summary assessment will be generated (such as in the case of synthetic image detection) by considering the results of multiple detectors into a single, user-friendly output. At the same time, users will retain the ability to access more granular details on demand, enabling deeper analysis when necessary.

Another significant development area is the integration of the model cards service developed in WP6 and from now on in WP7. In collaboration with the work in these WPs, MAAM will

incorporate structured model documentation that provides layered levels of transparency, from simplified explanations of each model's purpose and limitations to more technical descriptions including training datasets, performance metrics, and fairness, bias or factor analyses. This feature will improve trust and the interpretability of the results.

Enhancing the user feedback mechanism is also a key area of future work. Users will be able to rate detection outputs and provide textual feedback, supporting a more robust human-in-the-loop verification approach. In the current implementation, users can already contribute their own annotations, and the backend supports rating of AI-generated detection results. These capabilities will be further improved through a more user-friendly interface, making it easier for users to interact with and assess the outputs of the system. The commenting feature on assets, already implemented in the prototype, will also be improved and better linked to the verification tools, enabling discussions around specific detection results. This enhancement will facilitate collaborative media analysis, creating space for shared evaluations and insights. To further strengthen collaboration, MAAM will introduce an assignment mechanism that enables task distribution among workspace members. Additionally, workspace management will be upgraded to support more customized access rights, easier onboarding processes, and greater adaptability for large organizations. A user profile system will also be introduced, summarizing activity across the platform, such as past verification tasks, feedback contributions, and comments. This feature will help users to identify other experienced users within a collaborative workspace.

Another critical area of development is the ongoing improvement of the integrated annotation and detection models, in close coordination with the research activities in WP5. Given the fast evolution of generative AI and the risk of performance degradation due to new data distributions and manipulation techniques, MAAM must continuously evolve its detection models to maintain relevance and accuracy.

In addition, for deepfake detection, efforts will focus on integrating audio-based detection models to complement visual analysis, as this requirement emerged as a key need during the co-creation workshops conducted in WP3. In addition, the UI will be further developed with improved visualization tools that allow users to understand which parts of a video are flagged and why. As highlighted in the usage scenarios, there is a growing need for features that allow users to compare similar videos, especially those identified by the Near-Duplicate Detection (NDD) service. This will require enhancements including the ability to display videos side-by-side for direct comparison and validation. An important new feature will be the expansion of the platform beyond facial deepfakes to include detection of fully synthetic videos, which are becoming more prevalent with the advances in generative video models.

Last, in video analysis, annotation models will be enhanced to enable better object and person tracking, as well as improved automatic captioning, as the current method based on selected keyframes often leads to fragmented and incomplete annotation outputs. Future versions will aim for more continuous annotations throughout video sequences. To this end, we will further build on the video understanding approach described in Section 4.3.

These next steps will elevate MAAM into a **comprehensive, user-centered verification platform**, capable of adapting to emerging disinformation threats and serving the evolving needs of journalists, researchers, and fact-checkers. This will further be validated through the work in WP8, with the design and implementation of evaluation and validation methodologies.

7. References

- [1] Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., ... & Yang, H. (2022, June). Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning* (pp. 23318-23340). PMLR.
- [2] Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6202-6211).
- [3] Smaira, L., Carreira, J., Noland, E., Clancy, E., Wu, A., & Zisserman, A. (2020). A short note on the kinetics-700-2020 human action dataset. *arXiv preprint arXiv:2010.10864*.
- [4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
- [5] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016, September). Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision* (pp. 20-36). Cham: Springer International Publishing.
- [6] Zhang, Y., Huang, X., Ma, J., Li, Z., Luo, Z., Xie, Y., ... & Zhang, L. (2024). Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1724-1732).
- [7] Huang, X., Zhang, Y., Ma, J., Tian, W., Feng, R., Zhang, Y., ... & Zhang, L. (2023). Tag2text: Guiding vision-language model via image tagging. *arXiv preprint arXiv:2303.05657*.
- [8] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815-823).
- [9] Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018, May). Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)* (pp. 67-74). IEEE.
- [10] Sarriidis, I., Koutlis, C., Papadopoulou, O., & Papadopoulos, S. (2022, December). Leveraging large-scale multimedia datasets to refine content moderation models. In *2022 IEEE Eighth International Conference on Multimedia Big Data (BigMM)* (pp. 125-132). IEEE.
- [11] Koutlis, C., Schinas, M., & Papadopoulos, S. (2023). MemeTector: Enforcing deep focus for meme detection. *International Journal of Multimedia Information Retrieval*, 12(1), 11.
- [12] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.
- [13] Kordopatis-Zilos, G., Tzelepis, C., Papadopoulos, S., Kompatsiaris, I., & Patras, I. (2022). DnS: Distill-and-select for efficient and accurate video indexing and retrieval. *International Journal of Computer Vision*, 130(10), 2385-2407.
- [14] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
- [15] Podell, Dustin, et al. "Sdxl: Improving latent diffusion models for high-resolution image

- synthesis." arXiv preprint arXiv:2307.01952 (2023).
- [16] Mareen, H., Karageorgiou, D., Van Wallendael, G., Lambert, P., & Papadopoulos, S. (2024). TGIF: Text-Guided Inpainting Forgery Dataset. *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–6. doi:10.1109/WIFS61860.2024.10810690
 - [17] Guillaro, F., Cozzolino, D., Sud, A., Dufour, N., & Verdoliva, L. (2023, June). TruFor: Leveraging All-Round Clues for Trustworthy Image Forgery Detection and Localization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20606–20615.
 - [18] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *Journal of machine learning research* 21.140 (2020): 1-67.
 - [19] Chen, Xilun, et al. "VideoOFA: Two-Stage Pre-Training for Video-to-Text Generation." *arXiv preprint arXiv:2305.03204* (2023).
 - [20] Li, Bo, et al. "Llava-onevision: Easy visual task transfer." arXiv preprint arXiv:2408.03326 (2024).
 - [21] Chatti, Mohamed Amine, et al. "The LAVA model: Learning analytics meets visual analytics." *Adoption of data analytics in higher education learning and teaching* (2020): 71-93.
 - [22] Zhai, Xiaohua, et al. "Sigmoid loss for language image pre-training." *Proceedings of the IEEE/CVF international conference on computer vision*. 2023.
 - [23] Yang, An, et al. "Qwen3 Technical Report." arXiv preprint arXiv:2505.09388 (2025).
 - [24] Yu, Zhou, et al. "Activitynet-qa: A dataset for understanding complex web videos via question answering." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. No. 01. 2019.
 - [25] Patraucean, Viorica, et al. "Perception test: A diagnostic benchmark for multimodal video models." *Advances in Neural Information Processing Systems* 36 (2023): 42748–42761.
 - [26] Fu, Chaoyou, et al. "Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis." arXiv preprint arXiv:2405.21075 (2024).
 - [27] Kordopatis-Zilos, Giorgos, Symeon Papadopoulos, and Yiannis Kompatsiaris. "Geotagging social media content with a refined language modelling approach." *Intelligence and Security Informatics: Pacific Asia Workshop, PAISI 2015, Ho Chi Minh City, Vietnam, May 19, 2015. Proceedings*. Springer International Publishing, 2015.
 - [28] Koutlis, Christos, and Symeon Papadopoulos. "DiMoDif: Discourse Modality-information Differentiation for Audio-visual Deepfake Detection and Localization." *arXiv preprint arXiv:2411.10193* (2024).