EAGLE EYE

Multi Model Person Re-Identification

PROJECT SUPERVISOR

Dr. Muhammad Atif Tahir

PROJECT TEAM

Huzaifa Rashid (k21-3299)

Aarib Azfar (k21-3342)

Abdullah Ashar (k21-3189)

Submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science.

FAST SCHOOL OF COMPUTING

NATIONAL UNIVERSITY OF COMPUTER AND EMERGING SCIENCES KARACHI CAMPUS

May 2025

Project Supervisor	Dr. Muhammad Atif Tahir
Project Team	Huzaifa Rashid K213299 Aarib Azfar K213342 Abdullah Ashar K213189
Submission Date	May 15, 2025

Supervisor

Mr. Dr. Muhammd Atif Tahir

Head of Department

Dr. Ghufran Ahmed

FAST SCHOOL OF COMPUTING

NATIONAL UNIVERSITY OF COMPUTER AND EMERGING SCIENCES KARACHI CAMPUS

Acknowledgement:

All the thanks to the supervisor Dr. Atif Tahir and the members of the of the project Huzaifa Rashid, Aarib Azfar, Abdullah Ashar, and people who cooperated in providing the data for the project. All the mentioned people have contributed to the project and helped us on this journey. Without their help we could not have done this.

Table of Contents

ABSTRACT

Consider a system that can recognise a person based just on the phrase "a man in a red hoodie with a backpack." CLIP fails with fine-grained, instance-level reasoning that is necessary for text-based person retrieval, even if it offers a strong foundation for visual-text alignment. Inspired by IRRA, we expand on CLIP by introducing an Implicit Relation Reasoning module that learns cross-modal token connections through masked language modelling. In order to align image-text embeddings, we additionally use Similarity Distribution Matching (SDM), which minimises the KL divergence between their similarity distributions. Without the need for extra oversight or part annotations, our system provides a comprehensive end-to-end pipeline that enables users to add videos, text searches, and image targets, as well as retrieve and monitor results after recording. We show that our system is effective by achieving notable improvements on CUHK-PEDES.

Introduction

The goal of text-to-image person retrieval is to utilise a natural language description to identify a particular person in a big image gallery. This job, which offers useful applications ranging from public surveillance to personal media search, sits at the nexus of image-text retrieval and person re-identification (Re-ID). High intra-class visual variance and cross-modal differences between picture and text representations pose challenges to the task, despite its potential.

We use the IRRA (Implicit Relation Reasoning and Aligning) model to overcome these difficulties. This approach uses implicit local relation learning to improve global image-text alignment. In contrast to conventional global-matching or explicit local-matching techniques, IRRA uses self- and cross-attention to utilise fine-grained interactions between modalities without requiring extra processing at inference time. Our implementation employs a novel method in this context: masked language modelling (MLM) to direct the interplay between textual and visual characteristics during fine-tuning.

In addition, we suggest a novel training objective called the Similarity Distribution Matching (SDM) loss, which enhances cross-modal alignment by reducing the KL divergence between the ground truth and projected similarity distributions. To better highlight hard negatives, we have included a temperature parameter. We extensively fine-tune the entire CLIP model to maximise its pre-trained capabilities in multimodal representation learning, in contrast to earlier research that underutilise CLIP.

Our IRRA-based method achieves state-of-the-art performance with higher efficiency and discriminative power, as demonstrated by our evaluation on three benchmark datasets: CUHK-PEDES, RSTPReid, and FAST-NU.

RELATED WORK

We improved the performance of the preceding FYP group by optimising the fine-tuning of the CLIP-based dual-encoder to improve retrieval accuracy. In order to improve the results' interpretability and utility, we have added a novel function that allows the creation of detailed captions for recovered photographs. Our method achieves improved performance on benchmark datasets such as CUHK-PEDES, RSTP-ReID, and FAST University while retaining efficiency and utilising CLIP's strong cross-modal capabilities.

Additionally, Li et al. [6] first presented text-to-image person retrieval using the CUHK-PEDES dataset, which focusses on aligning text and image features in a joint embedding space for effective retrieval. Early techniques used matching losses for alignment and VGG and LSTM for feature extraction. This was further developed in later research by integrating better cross-modal matching losses for global feature alignment and by utilising ResNet50/101 [5] and BERT backbones. While some recent methods used attention processes for implicit local feature learning, others used human segmentation, body parts, and text phrases to add explicit local feature learning. Nevertheless, these techniques frequently make inference more computationally complex. Advanced vision-language pre-training models like CLIP were limited in their usage by the majority of earlier research' reliance on unimodal pre-trained backbones. Although CLIP was investigated for this job by Han et al. [4] and Yan et al. [8], they were unable to fully transfer its cross-modal alignment capabilities. Inspired by Transformer-based models such as BERT and ViT, Vision-Language Pre-training (VLP) is a powerful tool for learning multimodal representations from large-scale image-text pairs, which is useful for tasks such as visual question answering [1] and image captioning [2]. VLP models can be classified as dual-stream, which employ distinct encoders for quicker retrieval but lack sophisticated cross-modal interaction modelling, or single-stream [3], which concatenate modalities for a single transformer but are slower at inference.

Literature Review

Finding a person's photo from a gallery using a text description is known as text-to-image person retrieval, and it's becoming a more significant problem for applications like security and photo searches. Significant progress has been made recently, particularly with models that use vision-language pre-training, such as CLIP, to improve text and image alignment. By enhancing accuracy and incorporating a new function to produce captions for retrieved photos, our work expands on this and makes the results more comprehensible.

Using the IRRA model, which was trained on the CUHK-PEDES, RSTP-ReID, and FAST University datasets, we have improved accuracy by building on earlier work. Additionally, we have included a captioning tool that improves user usability by making retrieved photos easier to understand.

Methodology

1. Introduction to IRRA

With an emphasis on enhancing the alignment of text and picture features in a joint embedding space, the IRRA framework is put forth as a novel method for text-to-image person retrieval. It improves fine-grained interactions without the need for explicit local alignment, which can be computationally costly, by utilising the CLIP model, which has already been pre-trained on a large number of image-text pairs, and by introducing implicit relation reasoning. The system builds on earlier work from the user's FYP group, which concentrated on related retrieval tasks, and is trained on datasets like CUHK-PEDES, RSTP-ReID, and FAST University. The inclusion of a caption generation capability is a significant innovation that improves the usability and interpretability of the results that are obtained.

2. Feature Extraction Dual-Encoder

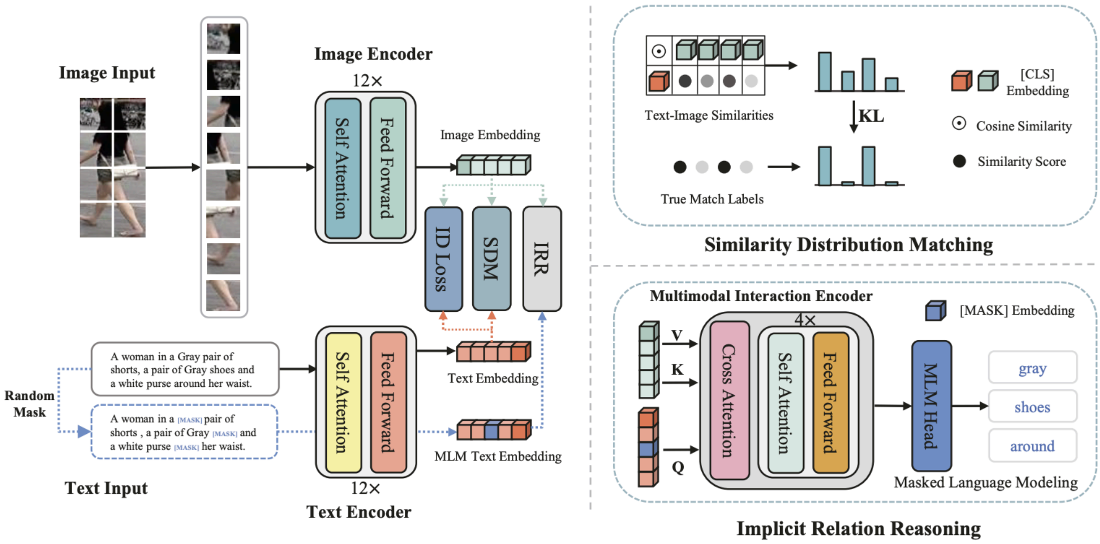
As demonstrated by Han et al. [4], IRRA starts with feature extraction using a dual-encoder architecture that was motivated by the partial success of transferring information from CLIP to text-image person retrieval. To improve underlying cross-modal alignment capabilities, IRRA initialises with the whole CLIP model immediately, in contrast to other research that usually utilise image and text encoders pre-trained separately on unimodal datasets.

3. Implicit Relation Reasoning (IRR)

Implicit Relation Reasoning, which focusses on implicitly mining fine-grained relations to learn discriminative global features, is introduced by IRRA in order to close the notable modality gap between visual and language. This is accomplished by using Masked Language Modelling (MLM), which was first put forth by Taylor [7] in 1953 and made popular by BERT. It has been modified for use in multimodal contexts

4. Similarity Distribution Matching (SDM)

Similarity Distribution Matching (SDM), a novel loss function introduced by IRRA, improves cross-modal alignment between text and images. This method uses KL divergence to align all image-text pairs inside a mini-batch while taking into account their similarity distribution. A softmax function is used to calculate each image's similarity to all text embeddings and transform that similarity into a probability distribution. Matching identity labels are used to generate a ground-truth distribution. The degree to which the anticipated similarity distribution resembles the ground truth is indicated by the SDM loss. To ensure mutual alignment, this procedure is used in both image-to-text and text-to-image directions. To avoid numerical instability, a little constant is included. Robust cross-modal representation learning is ensured by merging both directional losses to produce the final loss.



REQUIREMENTS

Functional requirements and the diagrams are given below:

USE CASES

Cropper:

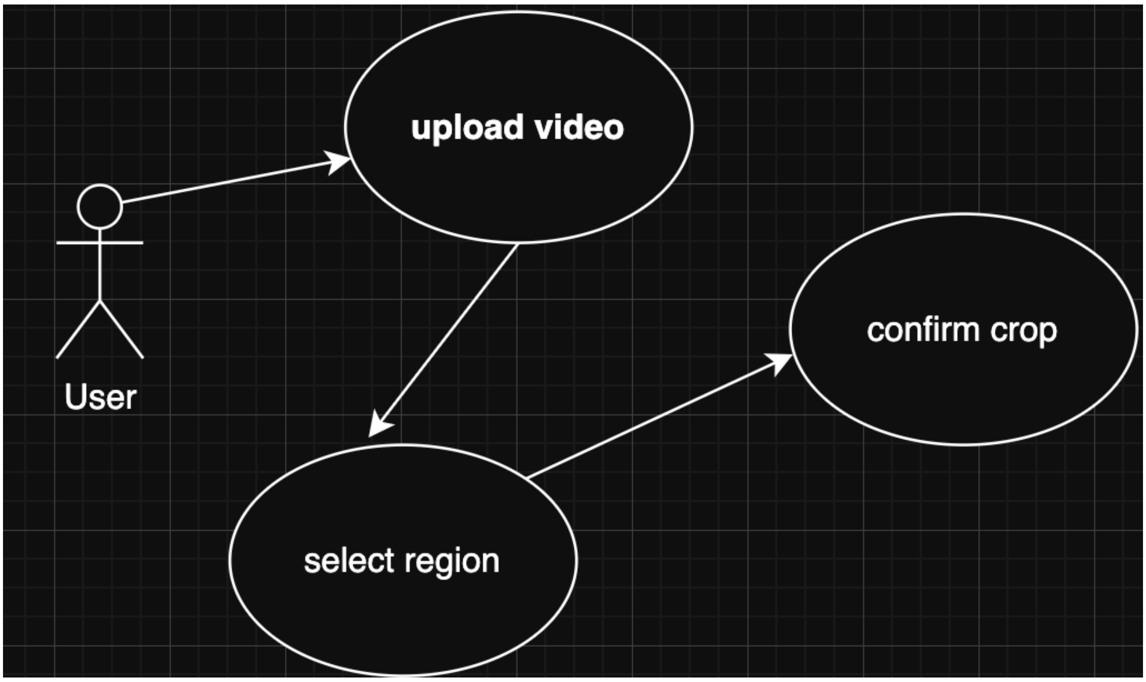


Figure 1: Use case Cropper

UC1 : Cropper	
Use case Id:	*Uc1*
Actors: *Users*	

Feature: *Cropper*		
Pre-condition:	*Video must be uploaded by user*	
Scenarios: *User wants to crop subjects from uploaded video*		
Step#	***Action***	***Software Reaction***
1.	*Upload a video*	*Video appears in the player*
2.	*Select timestamp and region*	*Cropping UI is shown*
3.	*Confirm crop*	*Cropped subject is saved to the gallery*
Post Conditions: *Cropped clips are available for re-ID tasks*		

Prompt Builder

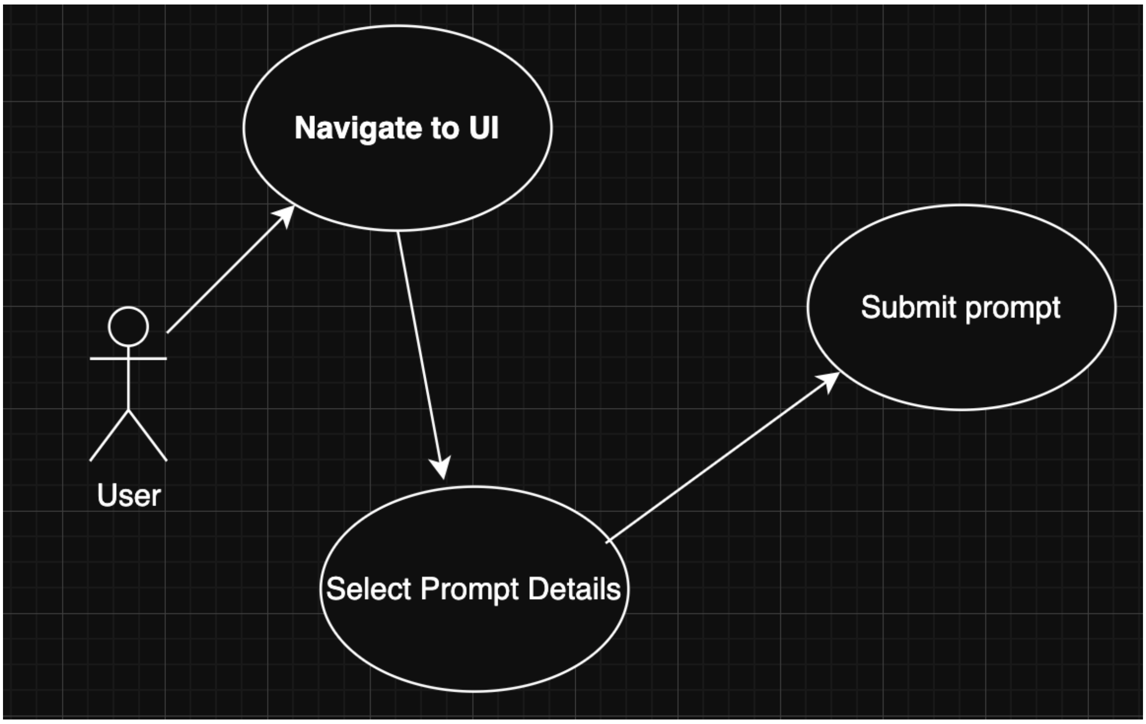


Figure 2: Prompt Builder Use case

UC2 : Prompt Builder		
Use case Id:	*Uc2*	
Actors: *User*		
Feature: *Prompt Builder*		
Pre-condition:	*User is logged in*	
Scenarios: *ser wants to generate detailed textual prompts*		
Step#	***Action***	***Software Reaction***
1.	*Navigate to Prompt UI*	*Prompt input box with tags is displayed*
2.	*Select prompt details*	*Prompt preview is generated*
3.	*Submit prompt*	*Prompt is stored and passed to inference API*
Post Conditions: *Successful authentication will lead the user to log into the system and use services.*		

Step#	Description
	Prompt is saved and visible in task summary
Use Case Cross referenced	Uc3 and Uc6

Textual Targets

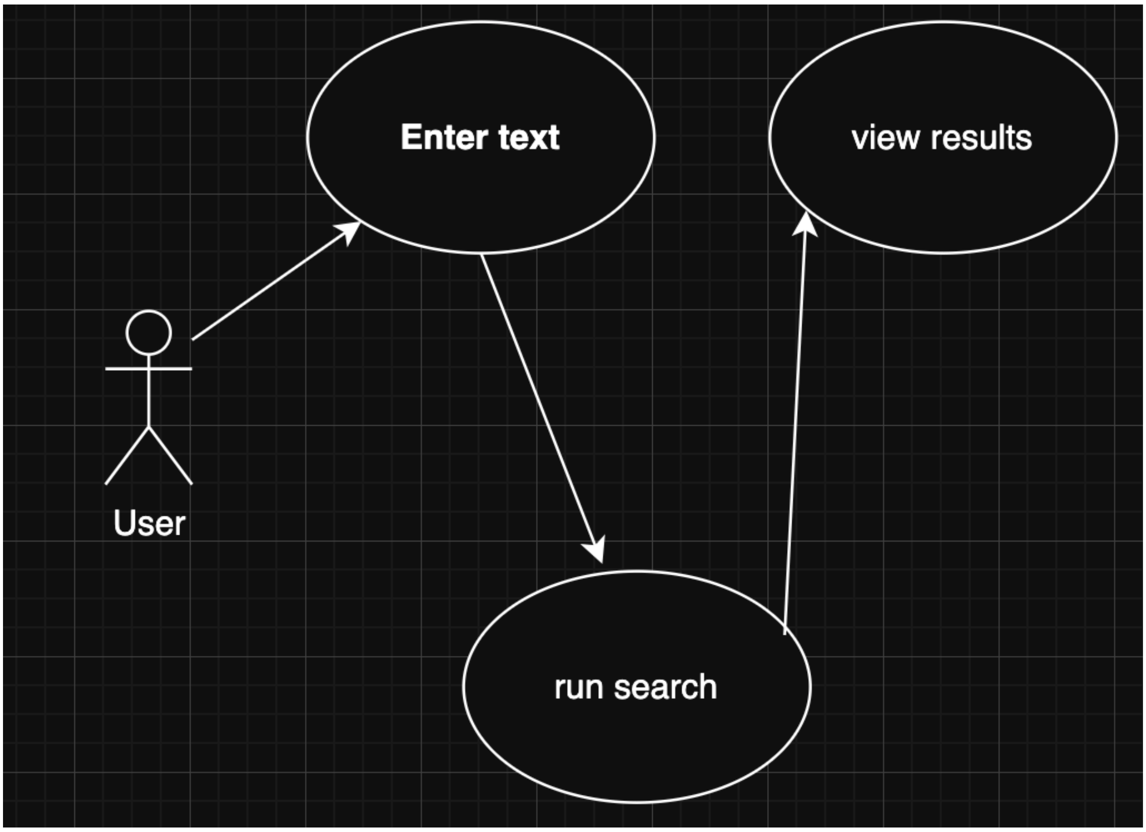


Figure 3: Textual Targets Use case

UC3: Textual Targets		
Use case Id:	Uc3	
Actors: User		
Feature: Textual Targets		
Pre-condition:	Target prompt must be created	
Scenarios: User wants to search with textual descriptions		
Step#	**Action**	**Software Reaction**
1.	Enter textual description	Backend parses and sends to CLIP model
2.	Run search	Matched subjects are returned and ranked
3.	View results	Results shown in dashboard
Post Conditions: Matches from video are highlighted		
Use Case Cross referenced	Uc1 and Uc2	

Visual Targets

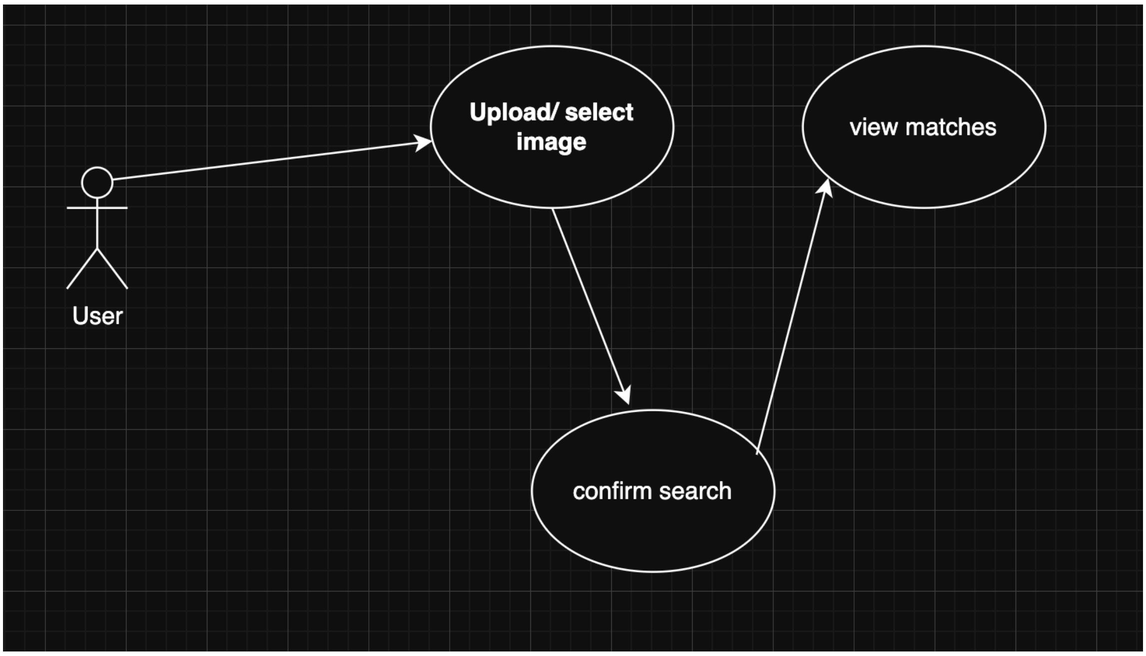


Figure 4: Visual Targets Use case

UC4: Visual Targets		
Use case Id:	*Uc4*	
Actors: *User*		
Feature: *Visual Targets*		
Pre-condition:	*Cropped visual or sketch uploaded*	
Scenarios: *User uses an image to find a target*		
Step#	***Action***	***Software Reaction***
1.	*Upload or select image*	*Image is processed via CLIP/ImageNet*
2.	*Confirm search*	*Results returned from inference*
3.	*View matches*	*Display of possible targets*
Post Conditions: *Results include visual matches*		
Use Case Cross referenced	*UC1, UC6*	

Gallery Management

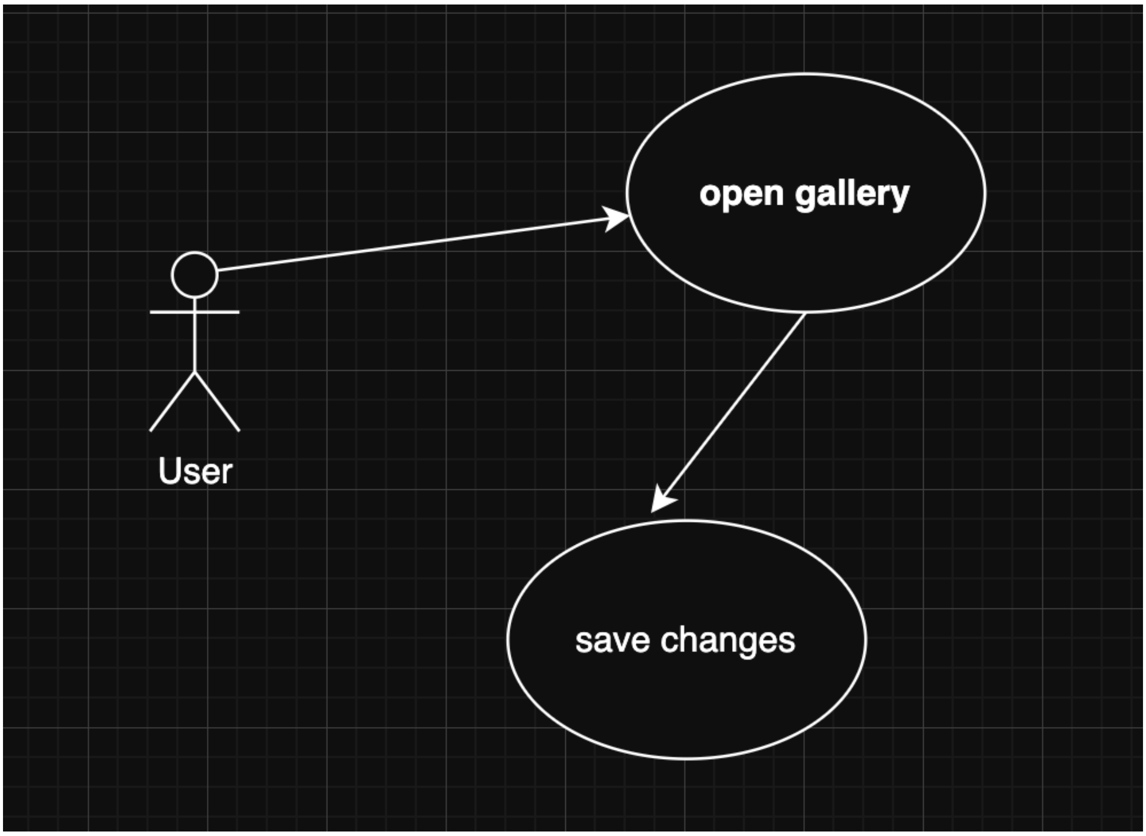


Figure 6: Gallery Management **Use case**

UC5: Gallery Management		
Use case Id:	*Uc5*	
Actors: *Users*		
Feature: *Gallery Management*		
Pre-condition:	*User has uploaded videos or crops*	
Scenarios: *User manages and groups sources*		
Step#	***Action***	***Software Reaction***
1.	*Open Gallery*	*LGallery view appears*
2.	*Save changes*	*Groups are stored in user profile*
Post Conditions: *View Services service providers.*		
Step#	***Description***	
1	*Video sources are grouped for future inference*	
Use Case Cross referenced	*UC1, UC6*	

Results Finetuning

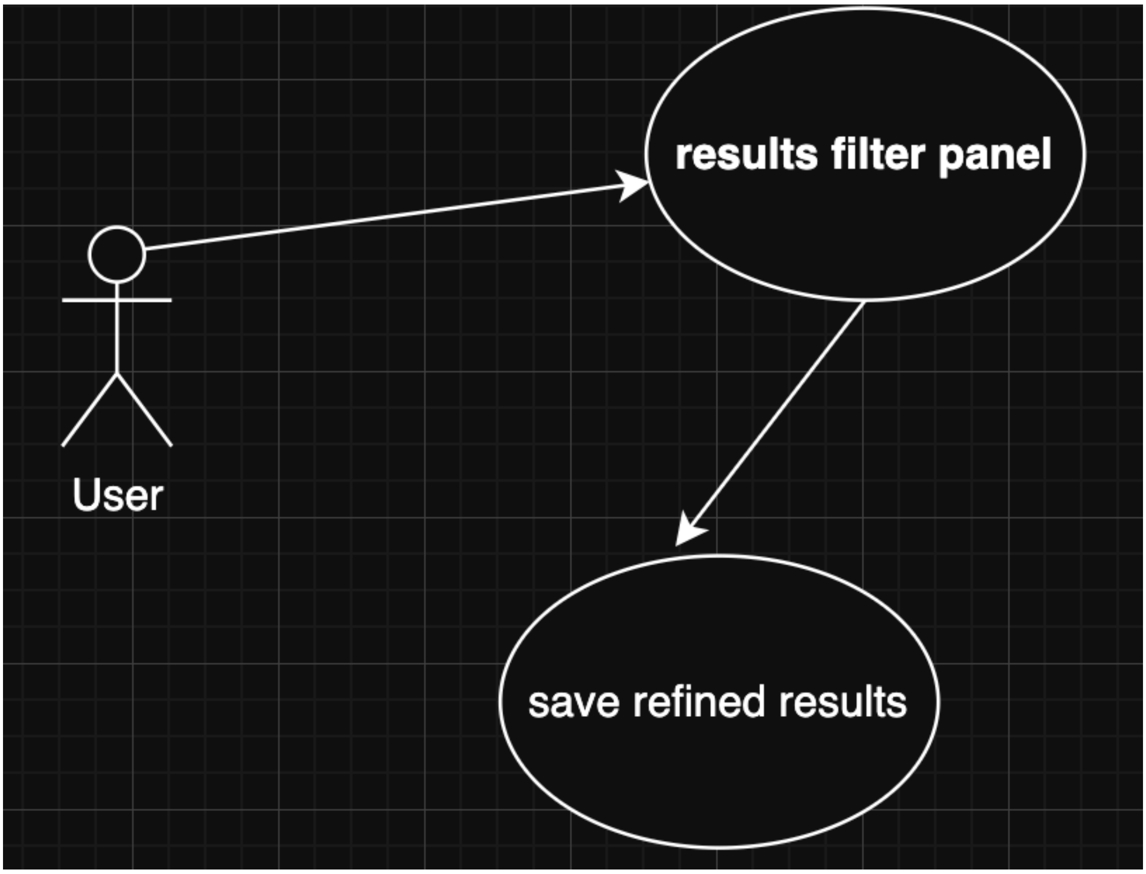


Figure 7: Results Finetuning Use case

UC6: Results Finetuning		
Use case Id:	*UC7*	
Actors:	*User*	
Feature:	*Results Finetuning*	
Pre-condition:	*Search must be completed with results available*	
Scenarios: *User wants to refine search results*		
Step#	***Action***	***Software Reaction***
1.	*Open Results Filter Panel*	*Filters and threshold controls shown*
2.	*Save refined results*	*Filtered set is saved to dashboard*
Post Conditions: *Final results reflect user-defined filters*		
Use Case Cross referenced	*UC3, UC4*	

DESIGN

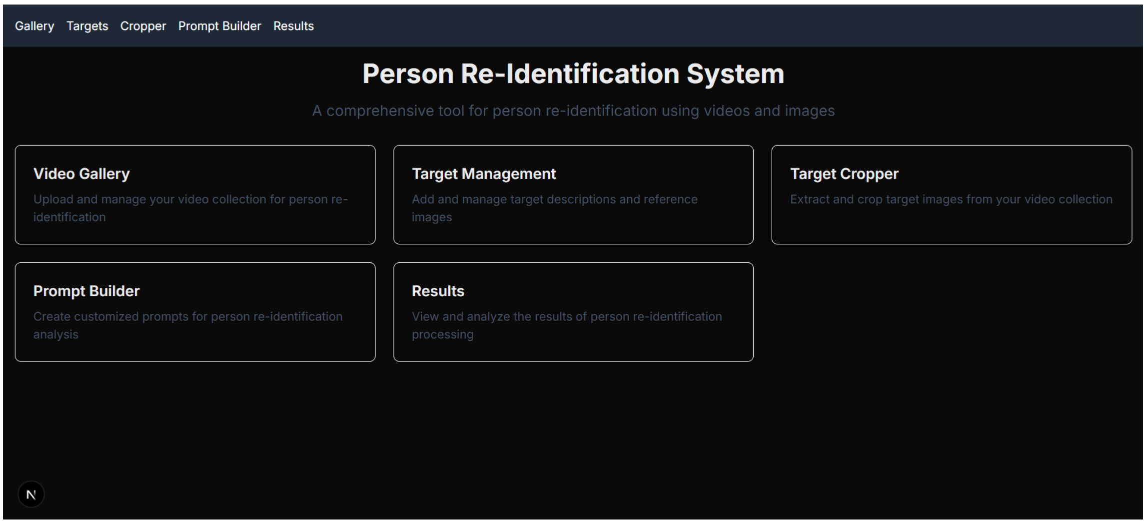
Eagle Eye uses Next.js to create a quick, responsive, and organised web application with an emphasis on usability and simplicity. The platform's main goals are to make it simple for users to submit video recordings, crop subjects, and conduct language and image-based searches. All of the main functions, such as Cropper, Prompt Builder, Textual and Visual Target Search, and Results Filtering, are designed to provide a seamless user experience.

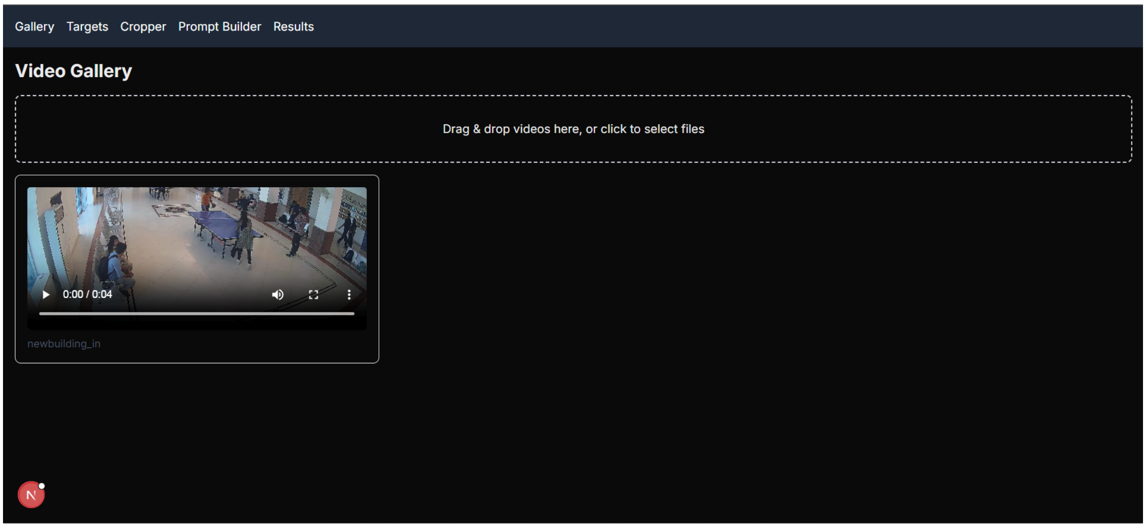
Users can interact with the system with ease thanks to the user interface's (UI) simple and straightforward layout. The overall experience is improved and user confusion is decreased by the clear buttons, forms, and interactions. Basic validation has been put in place to guarantee user input accuracy and avoid needless mistakes.

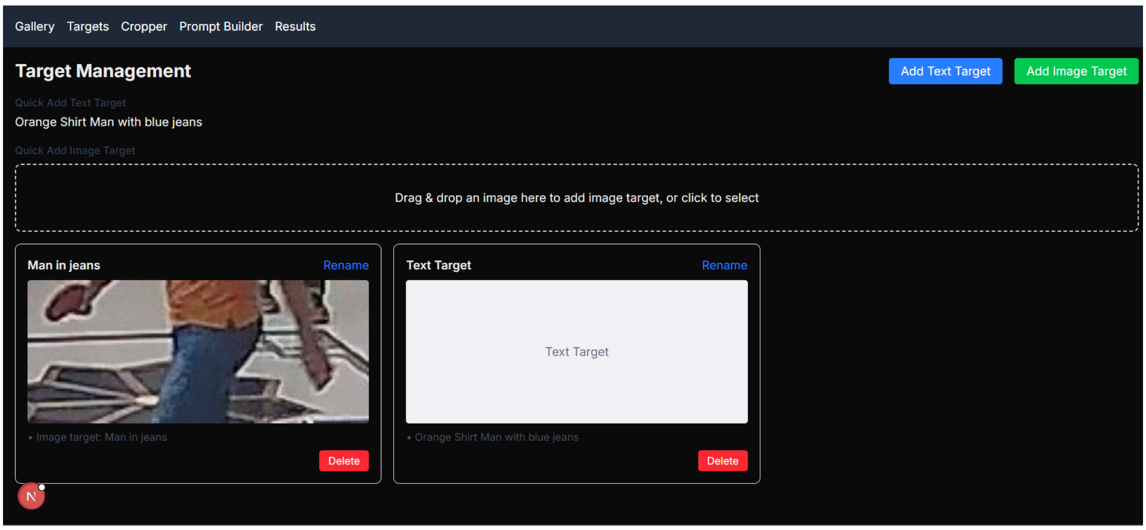
The system is set up on the backend to effectively manage search results, store prompts, and process video submissions. Despite having a straightforward structure, the platform guarantees that user actions and data flow are managed in a predictable and logical way.

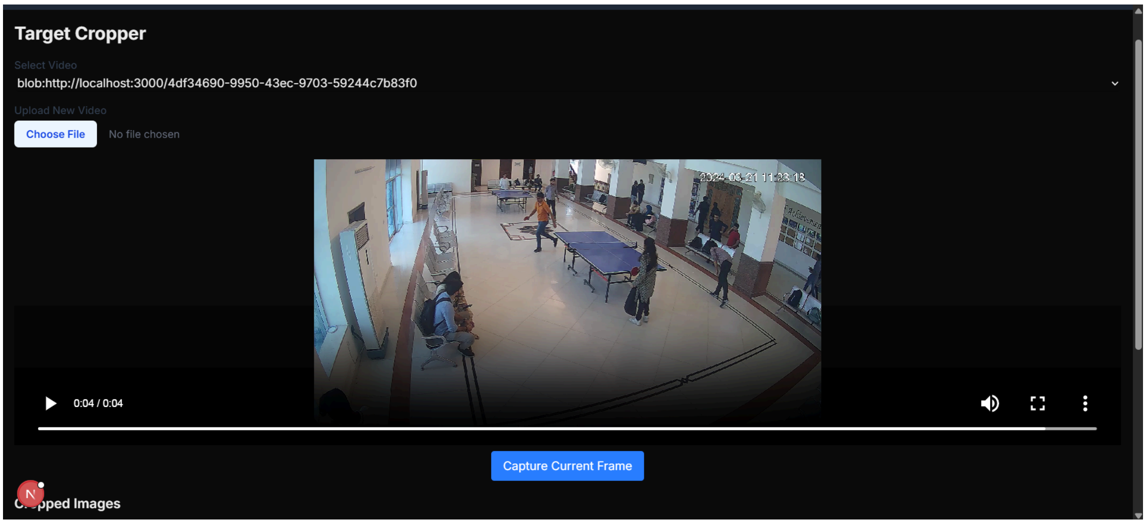
Our design strategy places a strong emphasis on usability, clarity, and ease of use so that users may maximise platform benefits without facing a challenging learning curve. Every interaction, whether trimming a topic or looking for a target, is designed to be quick and easy.

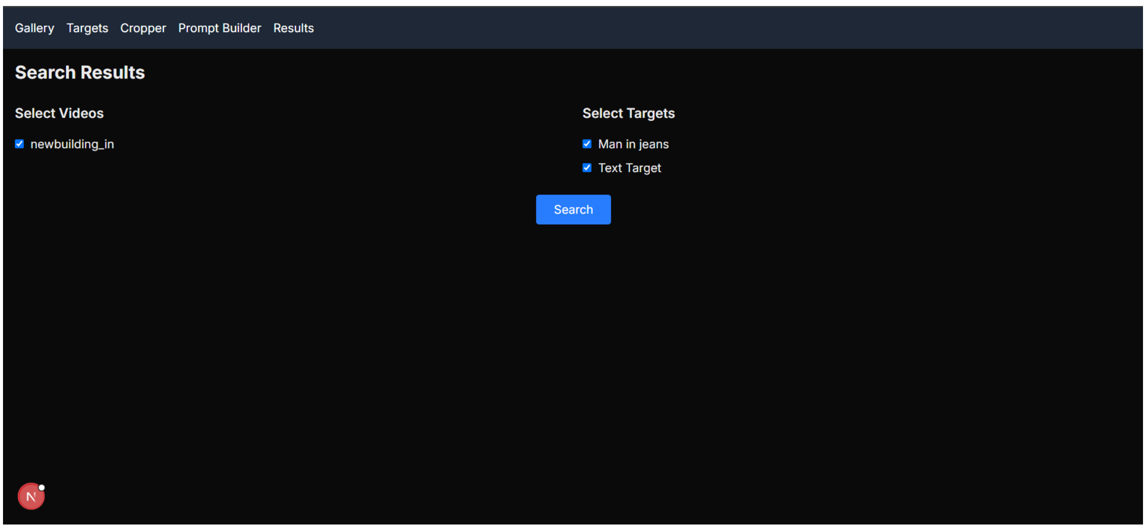
UserApp

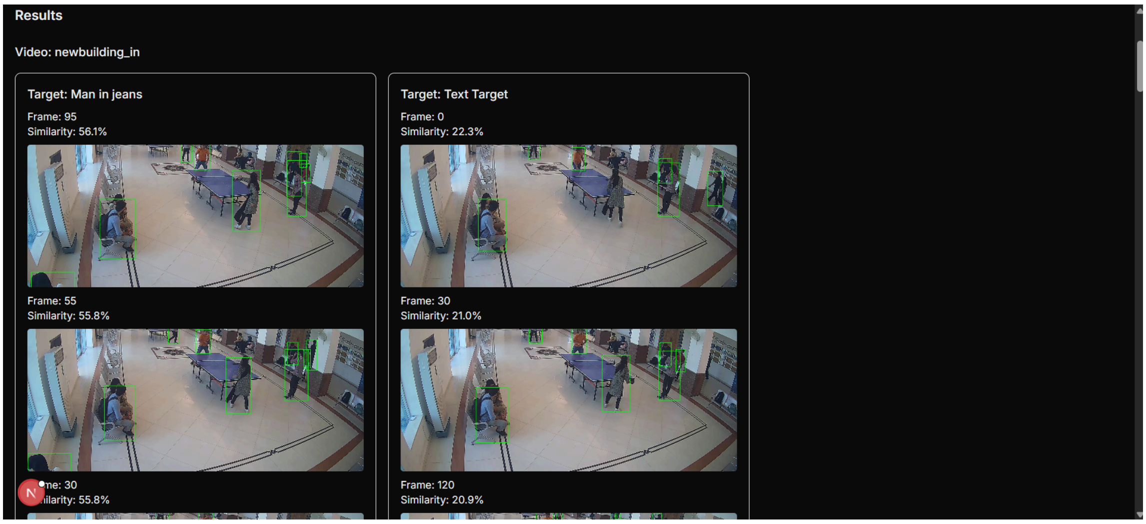

Gallery Targets Cropper Prompt Builder Results
Person Re-Identification System
A comprehensive tool for person re-identification using videos and images
Video Gallery
Upload and manage your video collection for person re-identification
Target Management
Add and manage target descriptions and reference images
Target Cropper
Extract and crop target images from your video collection
Prompt Builder
Create customized prompts for person re-identification analysis
Results
View and analyze the results of person re-identification processing


Gallery Targets Cropper Prompt Builder Results
Video Gallery
Drag & drop videos here, or click to select files
0:00 / 0:04
newbuilding_in


Gallery Targets Cropper Prompt Builder Results
Target Management
Add Text Target
Add Image Target
Quick Add Text Target
Orange Shirt Man with blue jeans
Quick Add Image Target
Drag & drop an image here to add image target, or click to select
Man in jeans
Rename
• Image target: Man in jeans
Delete
Text Target
Rename
Text Target
• Orange Shirt Man with blue jeans
Delete


Target Cropper
Select Video
blob:http://localhost:3000/4df34690-9950-43ec-9703-59244c7b83f0
Upload New Video
Choose File
No file chosen
0:04 / 0:04
Capture Current Frame
pped Images


Gallery
Targets
Cropper
Prompt Builder
Results
Search Results
Select Videos
newbuilding_in
Select Targets
Man in jeans
Text Target
Search



Results
Video: newbuilding_in
Target: Man in jeans
Frame: 95
Similarity: 56.1%
Frame: 55
Similarity: 55.8%
me: 30
ilarity: 55.8%
Target: Text Target
Frame: 0
Similarity: 22.3%
Frame: 30
Similarity: 21.0%
Frame: 120
Similarity: 20.9%

Implementation

The Eagle Eye system's implementation creates a smooth user experience for human re-identification activities by combining a Flask-based backend with a Next.js frontend. To submit video data, specify target users, and start inference procedures, users engage with the application via the frontend. The Flask backend manages the processing workflow by responding to API requests that are triggered by these activities. The inference subsystem processes video frames using the CLIP-based person re-identification model after receiving asynchronous inference jobs from the backend. These tasks are effectively scheduled and managed using a task queue. Processed video frames and identification outputs are examples of intermediate and final outcomes that are sent back to the backend before being sent to the frontend for user viewing. The complete system is implemented utilising cloud infrastructure such as EC2, ECS, or EKS, and all video files and extracted frames are kept in S3 or EFS storage solutions. Furthermore, CloudWatch is incorporated for performance monitoring and logging, guaranteeing system traceability and dependability. The services are containerised using Docker, which makes scalability and deployment easier.

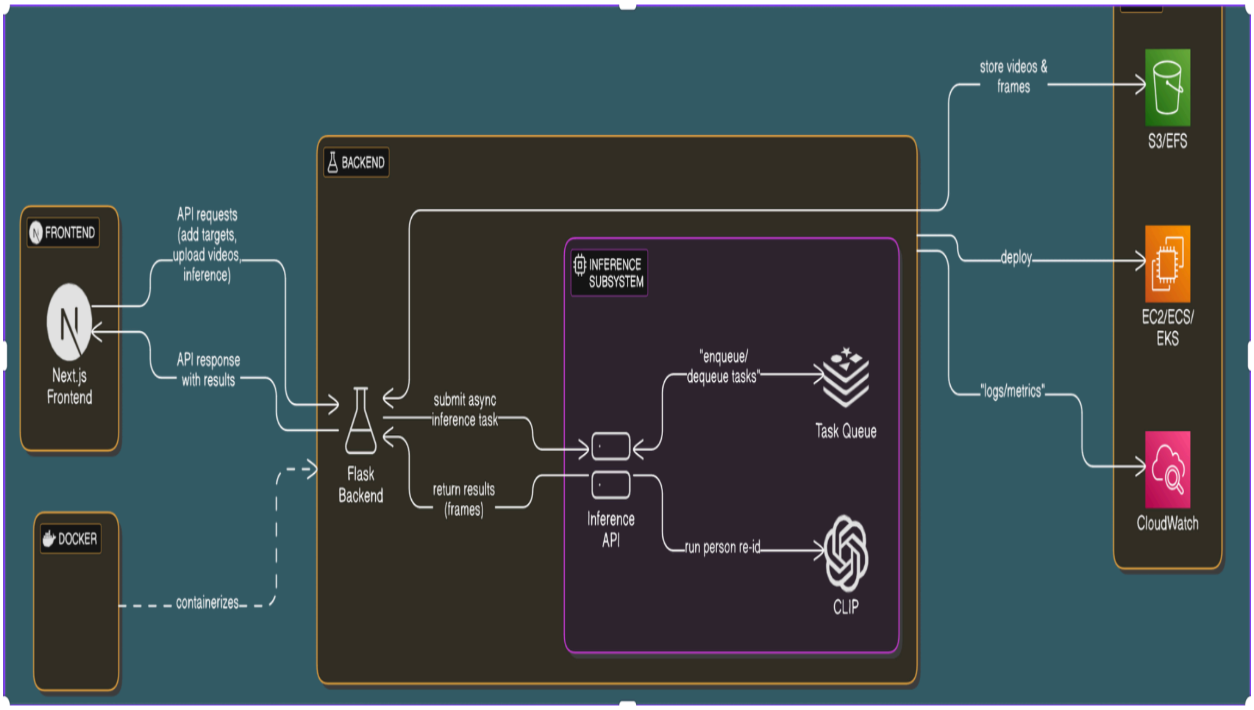


Figure 9: System architecture Diagram

TEST CASES

Test Cases	Names	Expected Result
TC1	User uploads a video through the frontend.	Video is successfully uploaded and acknowledged by the backend.
TC2	User adds target individuals via the UI.	*Target data is submitted and stored; confirmation message is shown to user.*
TC3	User submits an inference request.	System initiates the person re-identification task and returns a processing message.
TC4	Flask backend receives and processes the inference request.	Inference task is submitted to the task queue and CLIP model is invoked.
TC5	Inference subsystem dequeues and processes video frames.	Frames are processed and person identities are matched using the CLIP model.
TC6	Backend returns inference results to frontend.	Results are returned correctly and displayed to the user.
TC7	Check storage of video and frame data to S3/EFS.	All uploaded videos and generated frames are stored in the designated storage.
TC8	Verify deployment and system monitoring on AWS (EC2/EKS) with CloudWatch.	Logs and metrics are correctly sent to CloudWatch for system observability.
TC9	User receives real-time updates or task completion status on the frontend.	Frontend displays accurate status of submitted inference tasks.
TC10	Check error handling when uploading an unsupported file format.	User receives an appropriate error message and upload is rejected.

Table 1: Test cases

Conclusion

By combining textual, visual, and contextual inputs into a cohesive and intelligent system, Eagle Eye is a cutting-edge platform designed to enhance multimodal person re-identification. Eagle Eye's strong architecture, deep learning integration, and user-friendly interface enable analysts and security experts to reliably identify people in a variety of video sources and data modalities.

The platform tackles important issues in surveillance, monitoring, and forensic investigations by providing strong features like smart cropping, fast generation, visual/textual target detection, and gallery management. Eagle Eye's position as a game-changing tool in the security and analytics space is further supported by our dedication to providing accurate, comprehensible, and real-time findings. This solution not only enhances operational efficiency but also ensures a scalable and adaptive ecosystem, setting a new benchmark in the person re-identification landscape.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. *VQA: Visual Question Answering*. In Proceedings of the IEEE International Conference on Computer Vision, pages 2425–2433, 2015.

[2] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. *Microsoft COCO Captions: Data Collection and Evaluation Server*. arXiv preprint arXiv:1504.00325, 2015.

[3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. *UNITER: Universal Image-Text Representation Learning*. In European Conference on Computer Vision, pages 104–120. Springer, 2020.

[4] Xiao Han, Sen He, Li Zhang, and Tao Xiang. *Text-based Person Search with Limited Data*. arXiv preprint arXiv:2110.10807, 2021.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Deep Residual Learning for Image Recognition*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.

[6] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. *Person Search with Natural Language Description*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1970–1979, 2017.

[7] Wilson L. Taylor. *“Cloze Procedure”: A New Tool for Measuring Readability*. Journalism Quarterly, 30(4):415–433, 1953.

[8] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. *CLIP-driven Fine-Grained Text-Image Person Re-identification*. arXiv preprint arXiv:2210.10276, 2022.