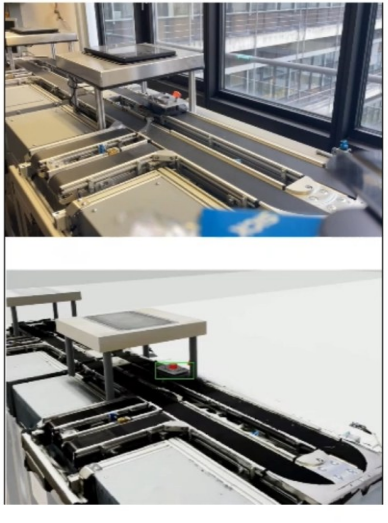
University of Stuttgart
Institute of Industrial Automation
and Software Engineering

real scenario



planned 3D scenario
(in simulation)

Vision-Language Model-Based Anomaly Detection and Analysis Using Simulated and Camera Data

Presenter: Jiawen Xu

Supervisor: Yuchen Xia M. Sc.

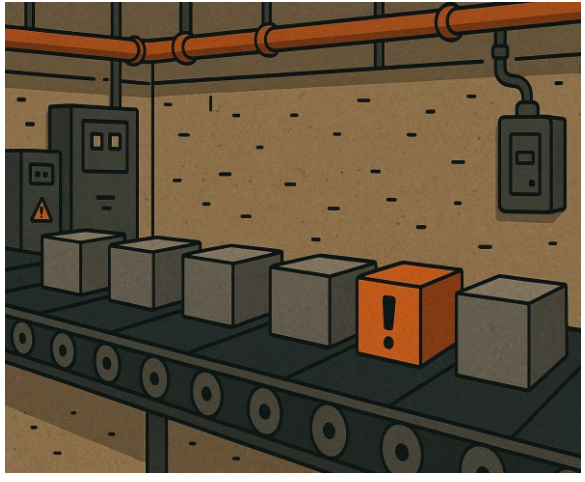
Examiner: Prof. Dr. Ing. Michael Weyrich

Introduction

- Use Case Scenario
- Problem Statement

Use case scenario

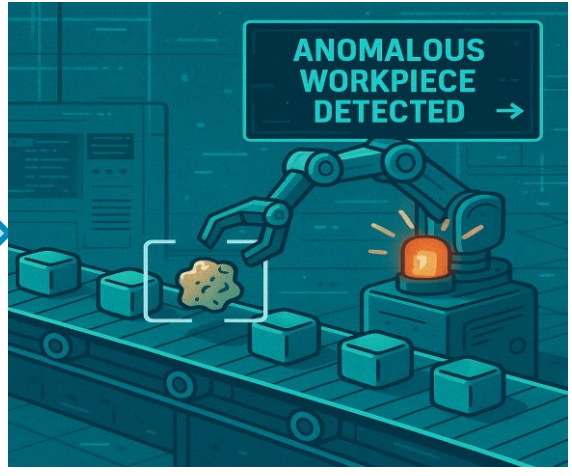
If an anomaly occasionally occurs on the production line..



An anomaly occurs in the production process

the operator detects it and reports it

manual



machine detects&reports

automated(data-driven)

- High labor cost: The process relies heavily on manual intervention.

Introduction

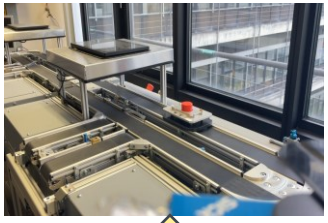
- Use Case Scenario
- Problem Statement

Problem Statement

Observed data (video)

HOW?

Video (taken from laboratory)



Object Recognition + Logics

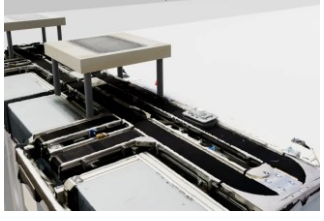
Comparative analysis

Anomaly?

Simulator

Simulation video

(NVIDIA ISAAC Simulator)
MT-3777



Reference data (video)

Basics

- State Of The Art
- Multimodal LLMs

State Of The Art

	Rule Based (~1980-~2005)	Machine Learning (~2005-~2018)	V-LLM (~2023-now)
Discription	Uses manually defined **rules** and **thresholds** for detecting objects based on features like **color**, **shape**, **size**, or **motion**.	Learns **patterns** from **labeled data** using models such as **SVMs**, Random Forests, or shallow **CNNs**	Uses pretrained multimodal models to understand **visual scenes** with **text prompts** or question
Strengths	• Simple • Interpretable • **no training needed**	• Better generalization • **learn from data**	• **highly flexible** • language-guided detection
Limitations	• Not generalizable • **sensitive to noise**	• **Requires labeled data** • feature engineering needed	• High computational cost • limited control
Examples	• Background subtraction[1] • color-based segmentation[2]	• PCB defect detection based YOLO v5 [3]	• GPT-4o [4] • Qwen2-VL [5]

[1] Stauffer, C., & Grimson, W. E. L. (1999). *Adaptive background mixture models for real-time tracking*.
[2] Shi, Jianbo, and Jitendra Malik. Normalized cuts and image segmentation.
[3] Darwish & Jain, "A Rule-Based Approach for Visual Pattern Inspection"
[4] https://openai.com/research/gpt-4o
[5] arXiv:2409.12191

State Of The Art

Rule Based CV method(~1980-~2005)

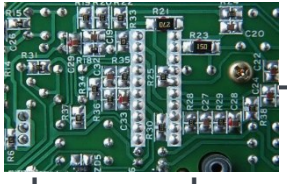
Adapted from: Darwish & Jain, "A Rule-Based Approach for Visual Pattern Inspection", IEEE TPAMI, 1988.

reference

Design rules, Model of reference shapes

defined rules

anomaly type	reference	observed
shape		
interruption		
color		
…	…	…



observed

Segmentation & Labeling

Each region is assigned a label (e.g., "pad," "resistor," "background")

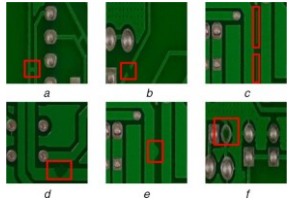
Feature Extraction

Area, shape, position…

Rule-Based Inspection

Is it in a valid position? Is the size within the expected range? Does the shape match?

Result: Pass / Defect Region

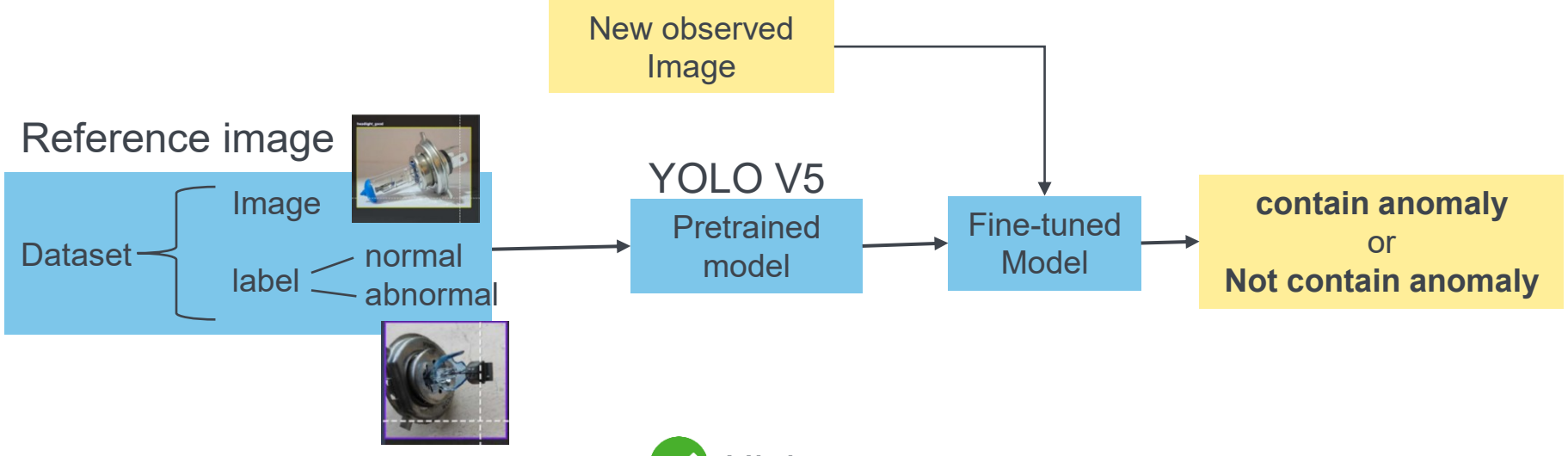


- Simple to implement
- Only suitable for static scnario
- Sensitive to noise and defects

State Of The Art

Machine Learning Model (**YOLO V5**)

Source: G. Rahull et al., "Automated Defect Detection System for Automobile Accessory Manufacturing Using YOLOv5," IEEE SCES, 2024.



- High accuracy
- Requires large dateset
- Limitations of predefined labels

Basics

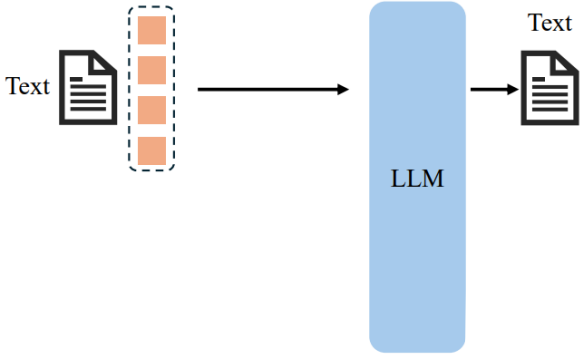
- State Of The Art
- Multimodal LLMs

Multimodal LLMs

Model	Type	Company	Features
GPT-4o	Proprietary	OpenAI	Benchmark reference, **High-quality, most popular**
Claude Sonnet	Proprietary	Google	**High-quality language output**
Qwen-2.5VL 7b	Open Source	Alibaba	**Lightweight**, low-cost
Qwen-2.5VL 32b	Open Source	Alibaba	Larger open-source model for vision precision comparison
Doubao Seed 1.6	Proprietary	Tik Tok	Low-cost, lightweight API-based model

Multimodal-LLMs

System Architecture

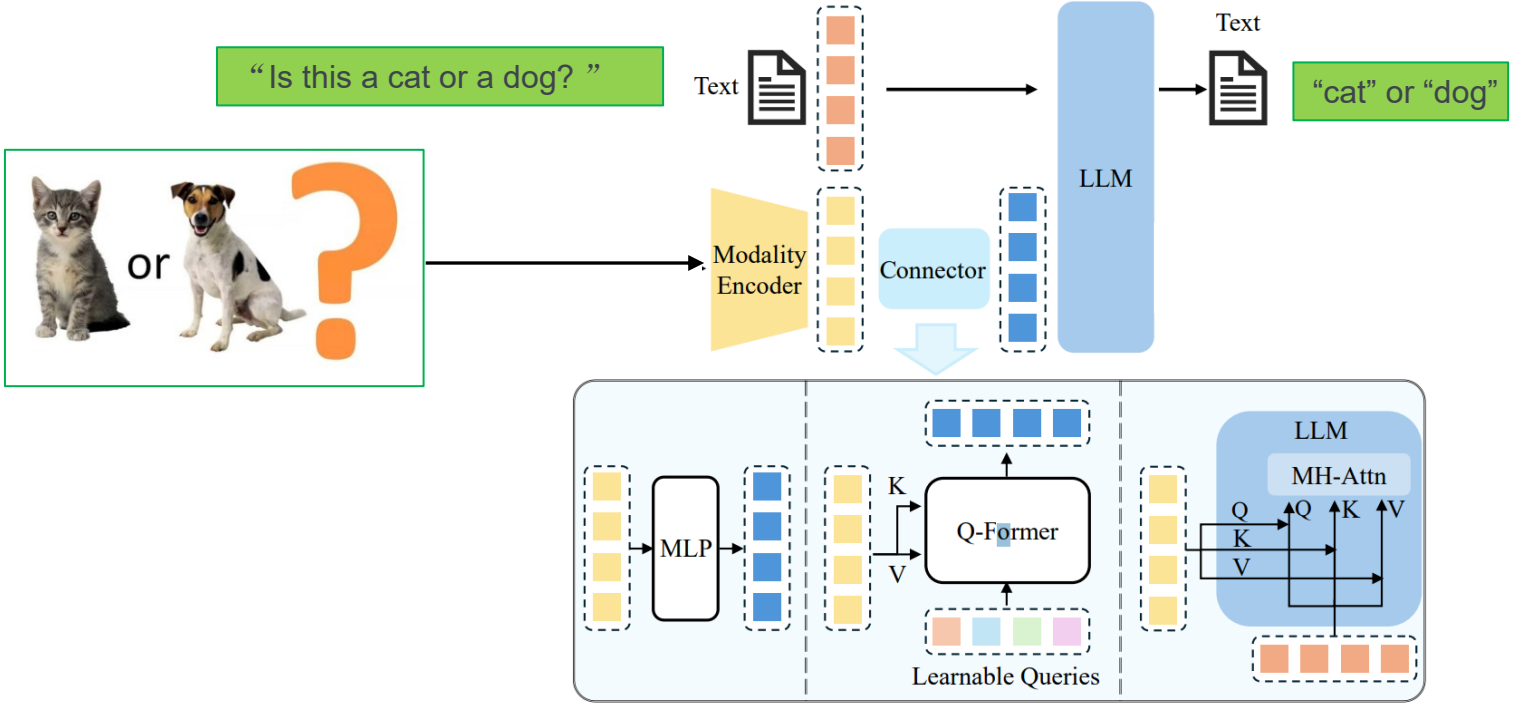


fuses visual and textual inputs

Source: Yin et al., A Survey on Multimodal Large Language Models, arXiv:2306.13549, 2023

Multimodal-LLMs

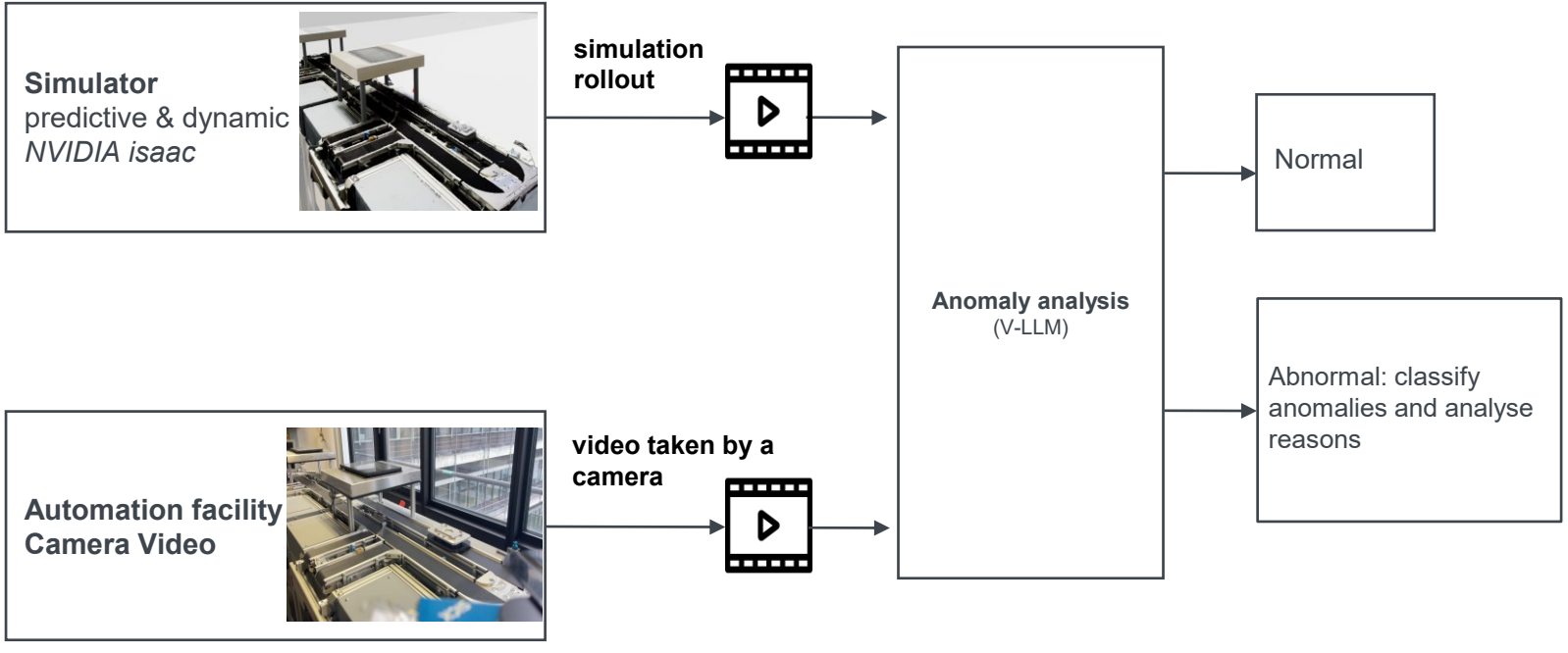
System Architecture



State Of The Art

V-LLM (this project)

Can V-LLMs be applied for industrial anomaly detection?

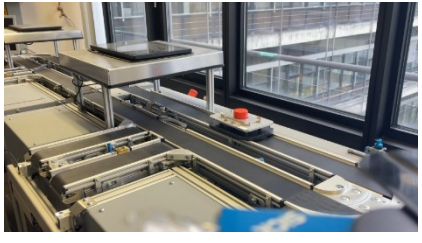


State Of The Art

V-LLM (this project)

It works!

Can V-LLMs be applied for industrial anomaly detection?



ChatGPT

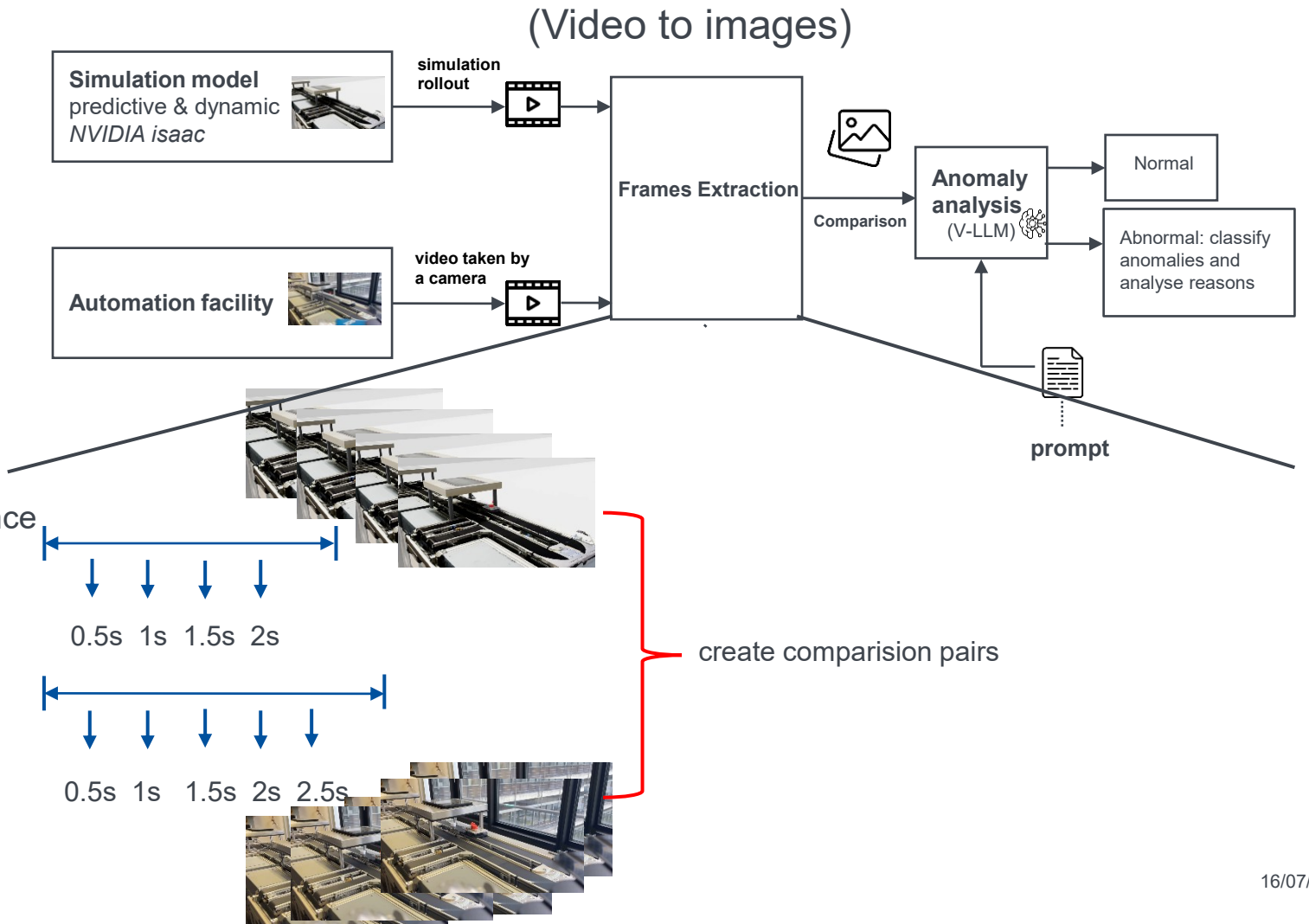
```
response = openai.ChatCompletion.create(
    model="gpt-4o",
    messages=[
        {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": "What are the main differences between these two images?"}
    ]
)
```

```
Yes, there are several noticeable differences between the two images:

1. Rendering Style:
   - The first image is a computer-generated 3D simulation or digital twin.
   - The second image is a real-life photograph of the same system.

2. Transported Object:
   - In the first image, the transport platform is empty.
   - In the second image, the transport platform is carrying a red emergency button.

3. Material and Surface Detail:
   - The real system in the second image shows more fine details (metal reflections, screws, wiring).
   - The simulation in the first image has simplified textures and fewer physical details.

4. Background:
   - The simulation background is plain white.
   - The real image shows a window view of a building exterior and part of a room interior.

5. Support Structure:
   - The platform support posts in the simulation appear as uniform grey cylinders.
   - In the real image, they are metallic with bolts and additional structural features.
```

System Design

- System Overview
- Dynamic ROI Detection

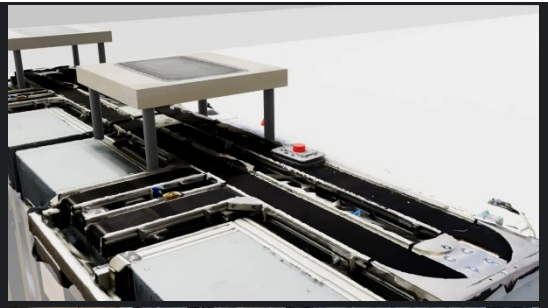
System Overview



System testing

Not reliable

```
### **Image 3:**
**Transport Object: Missing**
**Transport Platform: Present**
**Color: No red object detected**
**Path: Correct**
**Obstruction: None**
*Abnormality Detected: Transport object is missing*

### **Image 4:**
**Transport Object: Missing**
**Transport Platform: Present**
**Color: No red object present**
**Path: Correct**
**Obstruction: None**
*Abnormality Detected: Transport object is missing*

### **Final Judgement:**
*According to the criteria, both real-world frames show an abnormality due to the missing transport object.*

### **Summary (≤80 words):**
*The most likely causes are:*
**The transport object was not properly loaded onto the platform at the start of the process.**
**A detection or gripping failure during object placement resulted in an empty platform proceeding down the line.**
```

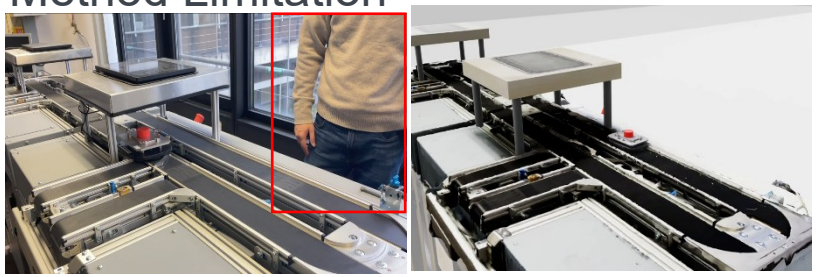


Expected output:
Foreign body on the conveyor

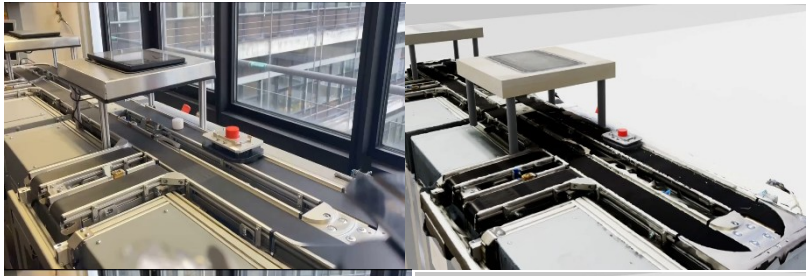
generated output:
The workpiece is missing

Test failures

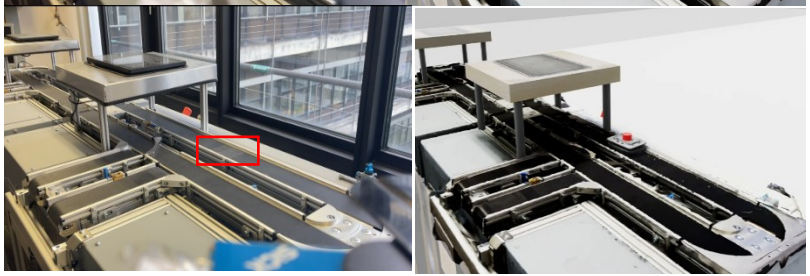
Method Limitation



Failure Type 1: environment object, pay false attention to enviroment



Failure Type 2: Different image visual rendering



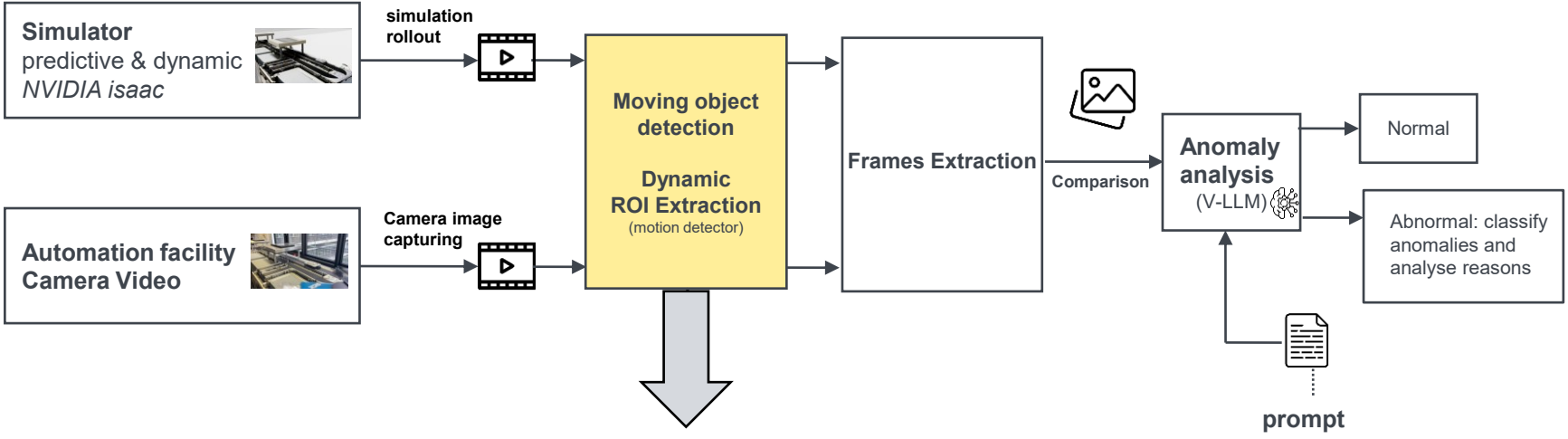
Failure Type 3: Sensitive to other noise (white dirt on the conveyor)

Summarized insight: → Model pays incorrect attention to noise

To improve the method: Make the model pay correct attention to the moving object!

System Overview

Method improvement



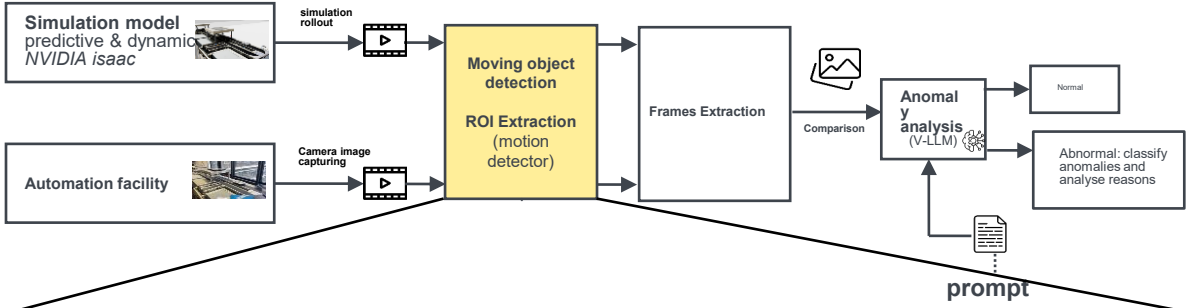
The model should pay more attention to the region containing moving object.

Called: Dynamic Region of Interest (Dynamic RoI)

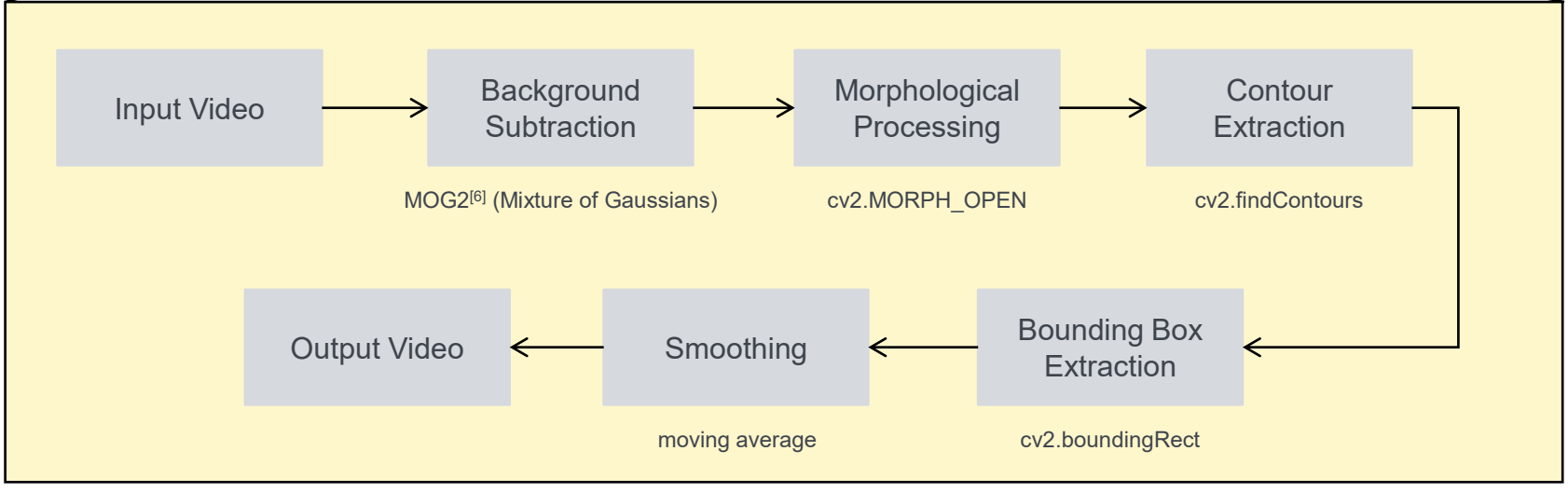
System Design

- System Overview
- Dynamic ROI Detection

Dynamic ROI Detection



Gaussian Mixture Model for Background Subtraction

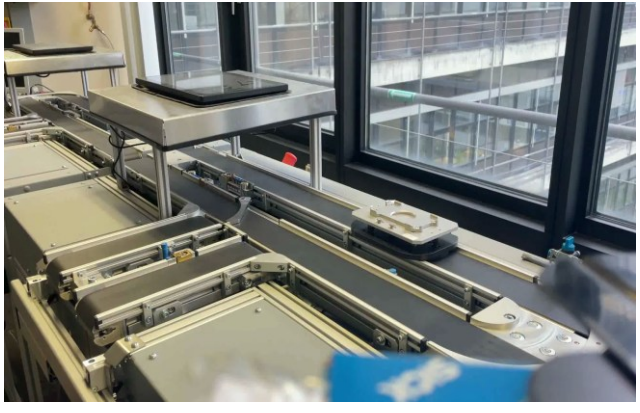


[6] MOG2: Z. Zivkovic, Improved Adaptive Gaussian Mixture Model for Background Subtraction, Proc. ICPR, vol. 2, pp. 28–31, 2004. doi: 10.1109/ICPR.2004.1333992

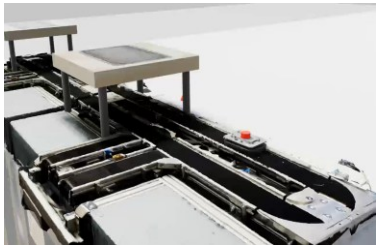
Dynamic ROI Detection - Result

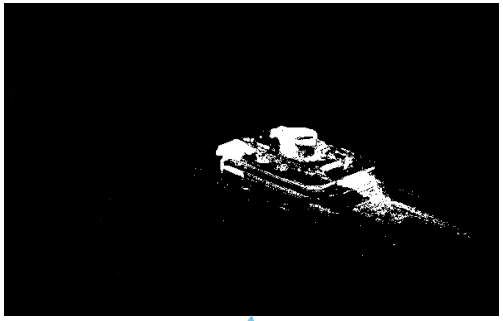
Raw Videos：

Videos annotated with ROI (in green box)



Dynamic ROI Detection



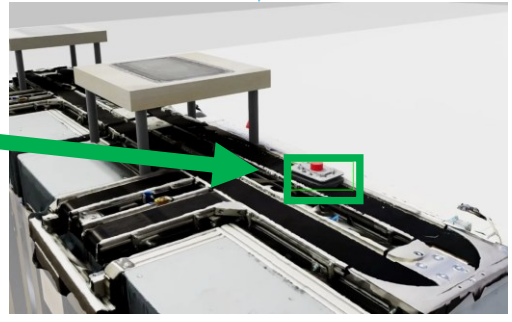


denoising

extract the object's boundary shape

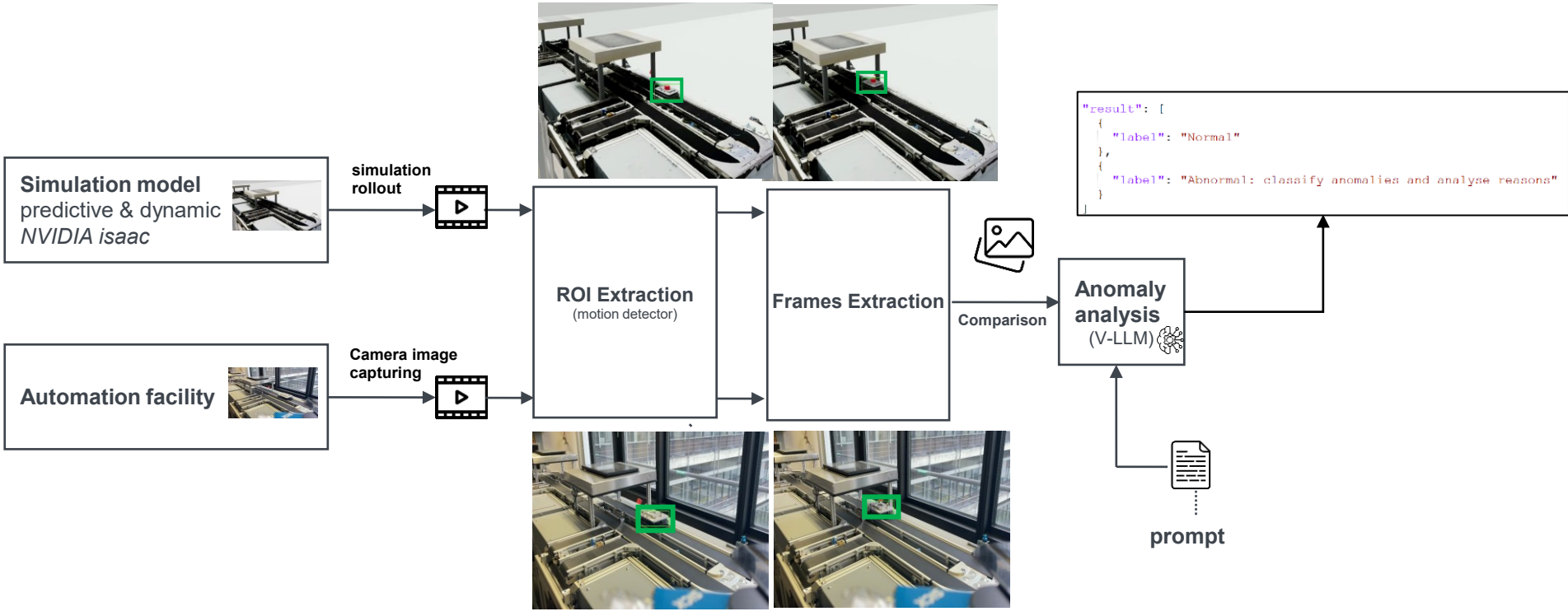
Input Video → Background Subtraction (MOG2 (Mixture of Gaussians)) → Morphological Processing (cv2.MORPH_OPEN) → Contour Extraction (cv2.findContours) → Bounding Box Extraction (cv2.boundingRect) → Smoothing suppresses frame-wise jitter (moving average) → Output Video

ROI



Optimized System

With dynamic object detection



Result – With Dynamic ROI Detection

```
### **Image 3:**
**Transport Object:** Present
**Transport Platform:** Present
**Object Color:** Incorrect – a white object is also present in front of the red object
**Path:** Correct
**Obstruction:** Yes – white object in front
*Abnormality Detected:* Presence of an unauthorized white object in front of the red object (object intrusion + color error)

### **Image 4:**
**Transport Object:** Present
**Transport Platform:** Present
**Object Color:** Incorrect – white object still present
**Path:** Correct
**Obstruction:** Yes – white object in front
*Abnormality Detected:* Same as above – unauthorized object intrusion (white) in front of the red object

### **Final Judgement:**
*Both real-world frames show abnormalities due to the presence of an additional white object in front of the expected red transport object.*

### **Summary (≤80 words):**
*The most likely causes are:*
**1.** An extraneous object (white) accidentally entered the conveyor system, causing a color error and obstruction.
**2.** A loading mechanism failure or human error allowed two objects onto the same platform.
```

Expected output:
Foreign body on the conveyor

generated output:
Presence of an unauthorized white object

Prompt Design

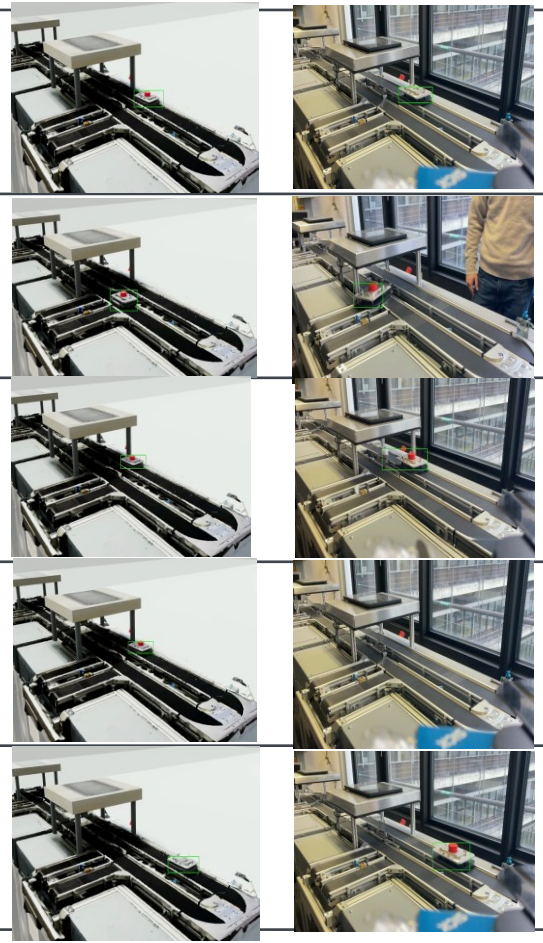
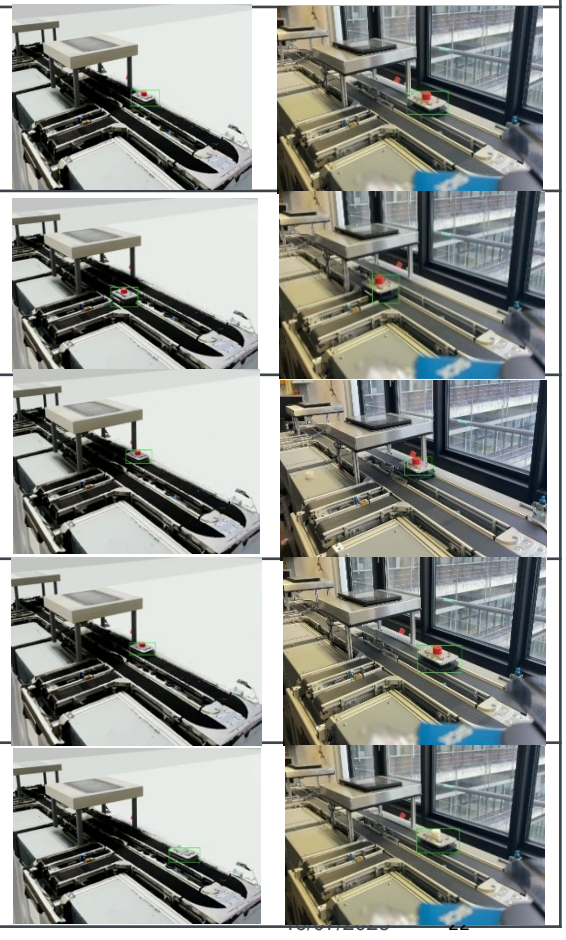
Prompt Section	Content
System Message	You are an industrial **image analysis expert**. I will provide you with multi-frame visual data, including a normal process (generated by simulation software) and a process (recorded from real-world). Analyze the process and identify any significant issues by comparing it to the normal one.
Instruction	Compare the first X **simulated frames** (normal) with the next X **real frames** (to analyze).
Image Input	[frames from simulated video] + [frames from camera video]
Follow-up Prompt	“ The process involves a conveyor belt that moves a transport platform. a red cylinder object is transported by the plantform on the conveyor belt. The transport path is controlled by a silver switch on the conveyor belt.” "If the path control switch malfunctions, it may lead to a routing error, causing the transport platform to go to the wrong workstation.” “ **Focus on the area marked by the green box in the image**.” “ Check only the following: (1) Is the transport object missing? (2) Is the transport plantform missing? (3) Is the object’s color correct? (4) if there is any other object in front of the transported object? (5) if the transport path is correct?” “ Note: The transport plantform is a distinct physical structure that moves on the conveyor belt and carries the object. Do not mistake conveyor belt shadows, reflections, or darker regions for the platform itself. ” “ Ignore people or components not located directly on the conveyor belt.” "Provide strictly factual, consistent, and reproducible answers without speculation or creative variation. Only focus on the given data, do not do uncertain assumption." “ Analyse frames from real-world video and determine whether these frames show any abnormalities, if show any abnomalities, return false and describe the type of anomaly and provide the two most likely causes for it within 80 words; otherwise only return true. “

Evaluation&Analysis

- Test cases
- Experiment Results
- Performance Comparison on Anomaly Detection：Precision & Recall

Test cases containing

1. Miss workpiece
2. Wrong path
3. Unexpected object appears
4. Miss transport platform and object
5. Wrong color/shape of the workpiece

Image pairs containing anomalies	(normal) image pairs that do **not** contain anomalies
	

Performance Comparison on Anomaly Detection：Precision & Recall

	Qwen 7b	Qwen 32b	chatGPT 4o	Doubao-Seed 1.6	Claude sonnet 4
normal	1 ✗ ✗ ✗ ✗ ✗ 2 ✓ ✓ ✓ ✓ ✓ 3 ✗ ✗ ✗ ✗ ✗ 4 ✗ ✗ ✗ ✗ ✗ 5 ✓ ✓ ✓ ✓ ✓	✗ ✗ ✗ ✗ ✗ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✗ ✗ ✗ ✗ ✗	✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✗ ✗ ✗ ✗ ✗ ✓ ✓ ✓ ✓ ✓	✓ ✓
abnormal	1 ✓ ✓ ✓ ✓ ✓ 2 ✓ ✓ ✓ ✓ ✓ 3 ✓ ✓ ✓ ✓ ✓ 4 ✓ ✓ ✓ ✓ ✓ 5 ✓ ✓ ✓ ✓ ✓	✓ ✓	✗ ✗ ✗ ✗ ✗ ✓	✗ ✗ ✗ ✗ ✗ ✓ ✓ ✓ ✓ ✓ ✗ ✗ ✗ ✗ ✗ ✗ ✗ ✗ ✗ ✗ ✓ ✓ ✓ ✓ ✓	✗ ✗ ✗ ✗ ✗ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✗ ✗ ✗ ✗ ✗ ✗ ✗ ✗ ✗ ✗

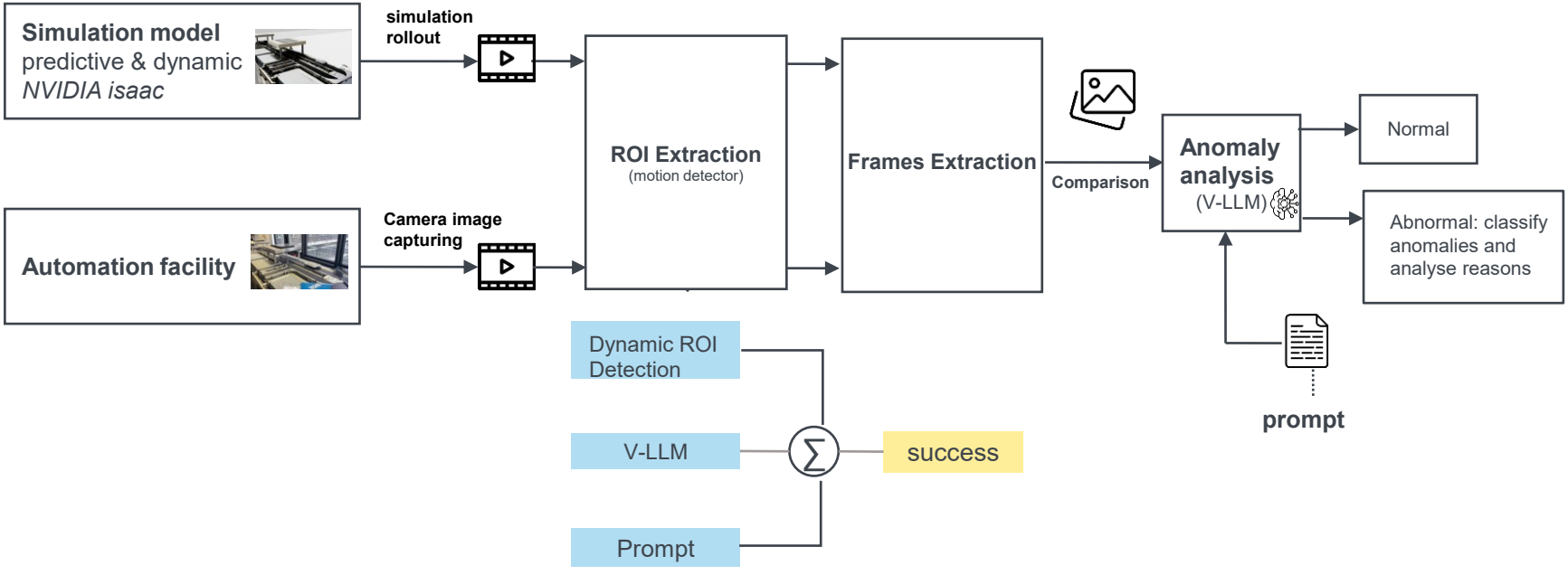
50 tests for 5 different v-LLM

Model	Precision	Recall	F1 Score
GPT-4o	100%	80%	**0.89**
Claude Sonnet	100%	40%	0.57
Qwen-2.5VL 7B	62.5%	100%	0.77
Qwen-2.5VL 32B	71.4%	100%	**0.83**
Doubao Seed 1.6	66.7%	40%	0.5

Conclusion

- Task completed & Future Work

Conclusion



Tasks Completed:

- simulation videos can serve as reliable references
- Multimodal large language models (V-LLMs) are capable of detecting anomalies

Future Work:

- Develop an automation software
- Expand testing to real-time industrial scenarios

University of Stuttgart
Institut of Industrial Automation
and Software Engineering

Thank you!

Jiawen Xu

e-mail st178020@stud.uni-stuttgart.de
phone +49 (0) 711 685-
fax +49 (0) 711 685-

University of Stuttgart
Institute of Automation and software systems