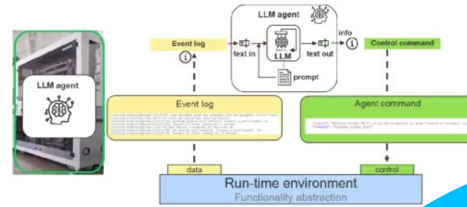
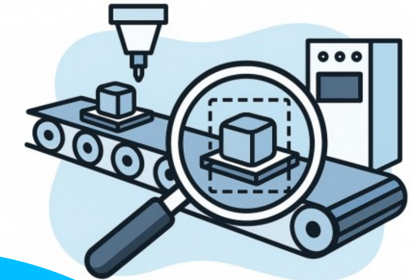
University of Stuttgart
Institute of Industrial Automation
and Software Engineering

LLM Agent
Generate command to control an automation module



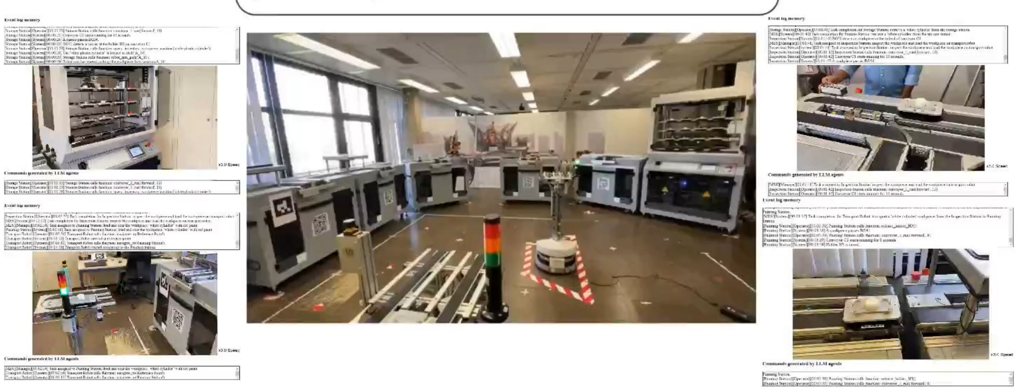


Integrating Vision-Language Models and Object Recognition for Image Analysis in Manufacturing Systems

Presenter: Zhongxin Cao
Supervisor: Yuchen Xia
Examiner: Prof. Dr. Ing. Michael Weyrich

Control Industrial Automation System with Large Language Models

User: “I need a coated drilled cylinder, please produce this for me.”

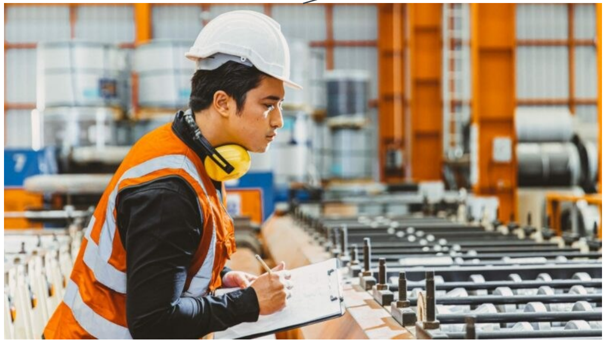


Y. Xia et. al,. “Control Industrial Automation System with Large Language Model Agents,” *https://arxiv.org/abs/2409.18009*

Contents

Application scenario

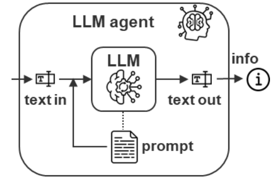
[00:00:25] Production plan is correctly executed.



Manual monitoring

VLM-driven dynamic monitoring

Previous Work



Event Log

[00:00:14] Sensor BG56 detects an object at the entrance.

sensor signal

control command

More perception of the process!

[00:00:27] Success. The workpiece A is placed on the carrier.
[00:00:29] Fail. The carrier is not positioned on conveyor C3.

LLM Generated Commands

[00:00:27] **[plan success]** The workpiece A is placed on the conveyor C3.

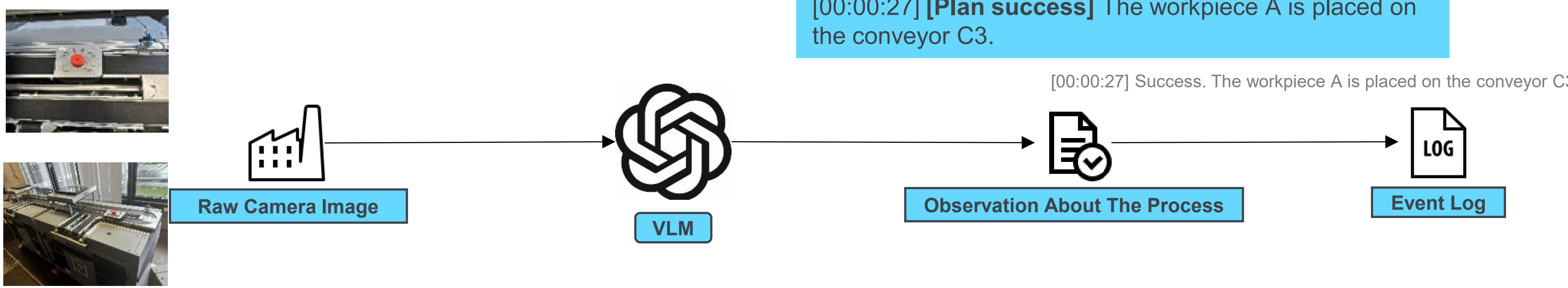
[00:00:29] **[plan failed]** The carrier is not positioned on conveyor C3.

Using VLM to perceive more information when a process is executed.

Intuition

More perception about the process:

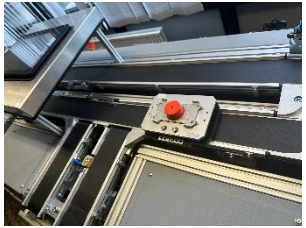
[00:00:27] **[Plan success]** The workpiece A is placed on the conveyor C3.



This conceptual design appears promising.

Therefore, a preliminary test was carried out to evaluate its practical effect.

Problem Statement



Production Scenario

```
Role
You are a professional industrial image analyst.

Input
One production-scene image (optionally with a bounding box around the target workpiece).

Requirements
1) Workpiece State
  • On the conveyor belt
  • (Optionally) on its designated carrier
  • In standby, awaiting processing

2) Conveyor Position & Movement (counterclockwise system)
  • Has just passed the station
  • Continues rightward along the belt
  • Is approaching the right turning part

3) Process Objectives (reference only; not for scoring)
  • Positional accuracy of the workpiece
  • Continuity of material flow

Evaluation Steps
1) Verify Workpiece State (belt / carrier / standby).
2) Verify Position & Movement relative to the station and rightward motion toward the right turning part (under counterclockwise flow).
3) Decision Rule: Output "Success" only if all required items in (1) and (2) are satisfied from the image; otherwise output "Fail".
4) If "Fail", give one brief, position-related reason (e.g., "not on carrier", "still before station", "moving leftward", "not approaching right turn").

Answer Format (strict JSON)
Return only the JSON below (no extra text, no trailing commas):
{
 "Result of planned process": "Success or Fail",
 "Description": "one brief explanation (position-related)"
}
```

Prompt

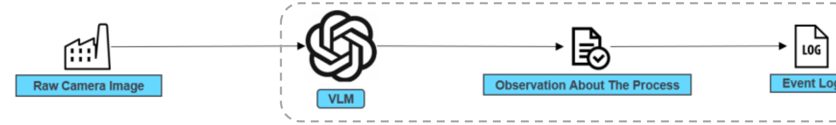
```
{
  "Result of planned process": "Success",
  "Description": "Workpiece is on carrier, moving rightward after station toward right turning part"
}
```

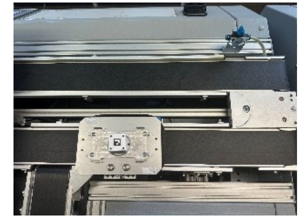
Output

```
[00:00:29] Success. Workpiece is on its carrier, positioned on the conveyor, moving rightward toward the turning section.
```

Event log

Success rate: 66.7% (16/24)





```
Role
You are a professional industrial image analyst.

Input
One production-scene image (optionally with a bounding box around the target workpiece).

Requirements
1) Workpiece State
  • On the conveyor belt
  • (Optionally) on its designated carrier
  • In standby, awaiting processing

2) Conveyor Position & Movement (counterclockwise system)
  • Continues rightward along the belt

3) Process Objectives (reference only; not for scoring)
  • Positional accuracy of the workpiece
  • Continuity of material flow

4) If "Fail", give one brief, position-related reason.
  • If a C2/C3 position comparison is possible, explicitly mention it (e.g., "workpiece still on C2, not on C3").
  • If not, provide a simple position-related reason (e.g., "not on carrier", "before station", "moving leftward").

Evaluation Steps
1) Verify Workpiece State (belt / carrier / standby).
2) Verify Position & Movement relative to the station and rightward motion toward the right turning part (under counterclockwise flow).
3) Decision Rule: Output "Success" only if all required items in (1) and (2) are satisfied from the image; otherwise output "Fail".
4) If "Fail", give one brief, position-related reason.
  • If a C2/C3 position comparison is possible, explicitly mention it (e.g., "workpiece still on C2, not on C3").
  • If not, provide a simple position-related reason (e.g., "not on carrier", "before station", "moving leftward").

Answer Format (strict JSON)
Return only one JSON object in the exact format specified (no extra text, no explanation):
{
 "Result of planned process": "Success or Fail",
 "Description": "one brief explanation (position-related)"
}
```

```
{
    "Result of planned process": "Fail",
    "Description": "No workpiece detected on the conveyor belt."
}
```

```
[00:00:30] Fail. No workpiece detected on the conveyor belt.
```

Not reliable

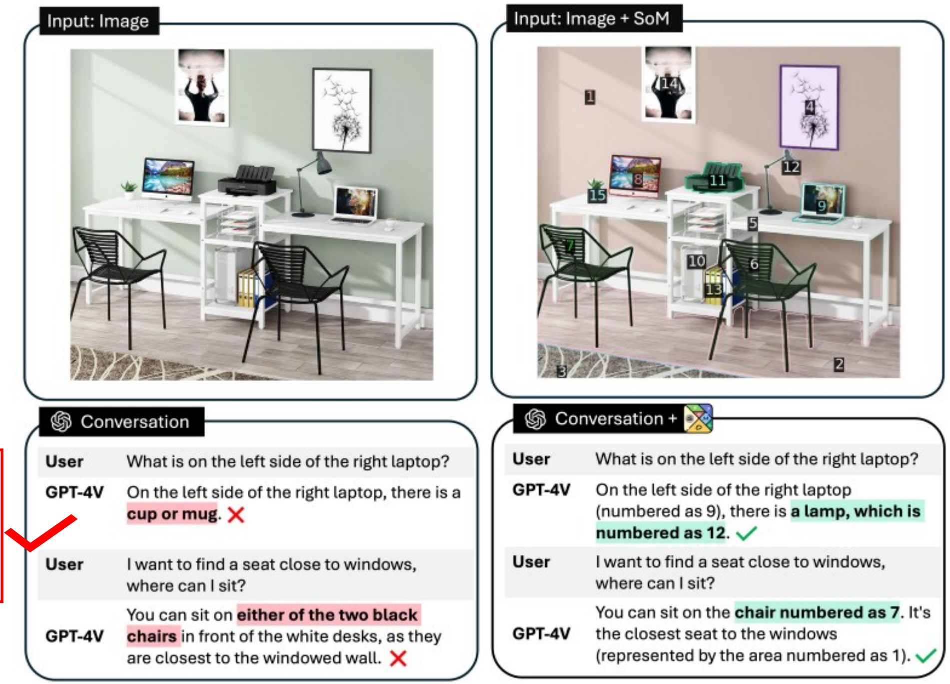
Key reason: **Visual context misinterpretation**

Literature Review

Solve this problem with image context annotation.

Annotation

Method	Strengths	Weaknesses
Graph-Based Context Reasoning [2]	Suppresses unlikely detections, improves precision	Weaknesses: Relies on predefined context; limited for rare/unseen objects.
Hierarchical Multi-Level Context [3]	Robust in cluttered scenes	May be misled by global context; computationally expensive
Multi-Image Collage Prompting [4]	Cost-efficient; improves multi-image understanding	Layout-sensitive; risk of confusion between adjacent images
Set-of-Mark Prompting [5]	**Zero-shot fine-grained grounding; state-of-the-art performance**	Depends on segmentation quality; best for explicit spatial reasoning tasks



[5]

[2] J. Ji, K. Ye, Q. Wan, and L. Shen, "Reasonable Object Detection Guided by Knowledge of Global Context and Category Relationship," *Expert Systems with Applications*, vol. 209, Article 118285, 2022.
[3] Z.-M. Chen, X. Jin, B. Zhao, X.-S. Wei, and Y. Guo, "Hierarchical Context Embedding for Region-based Object Detection," *ECCV*, pp. 423–439, 2020.
[4] S. Xu, Y. Wang, D. Liu, B. Du, and C. Xu, "CollagePrompt: A Benchmark for Budget-Friendly Visual Recognition with GPT-4V," *NAACL* 2025, pp. 6396–6418, 2025.
[5] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao, "Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V," *arXiv* preprint arXiv:2310.11441, 2023.

Enhancing Image Context — Masks and Annotation

Paper[5]: Annotation Marks → better VLM understanding

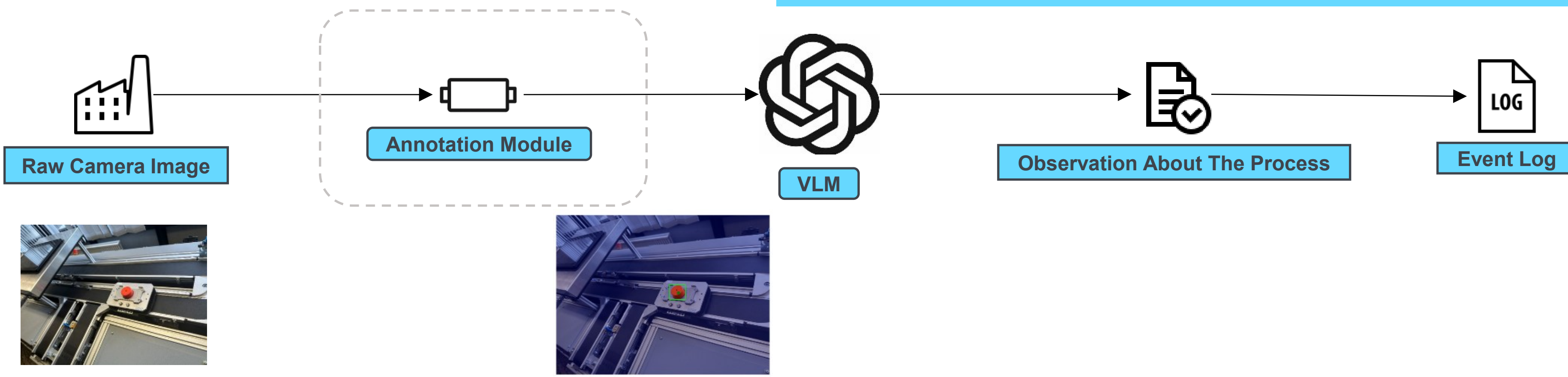
Annotate these objects

Annotation → **More focused visual context**

How to annotate image context?

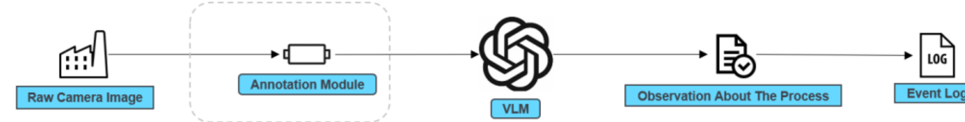
→ This naturally leads to **object recognition methods**, which we will compare in the next section.

[00:00:27] **[plan success]** The workpiece A is placed on the conveyor C3..



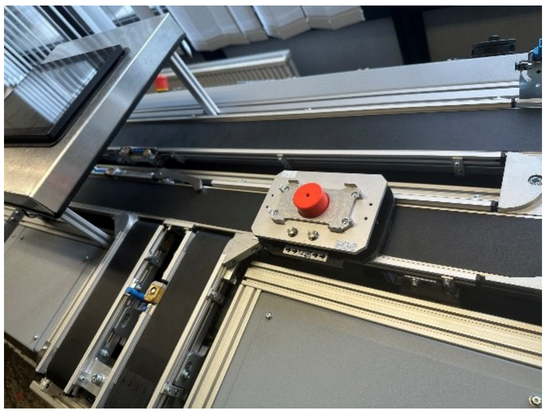
[5] Y. Zhang, J. Li, X. Wang, et al., "Set-of-Mark prompting unleashes extraordinary visual grounding in GPT-4V," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

Contribution

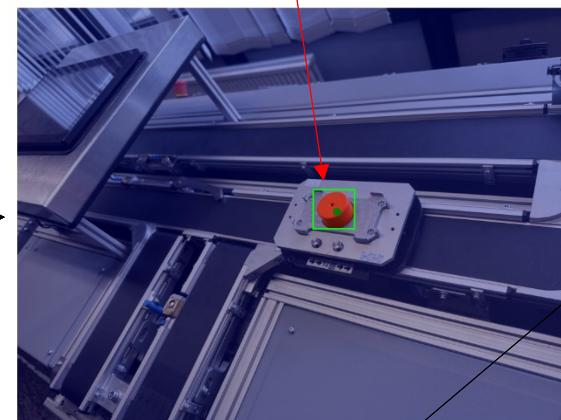


Solve image annotation task with object recognition methods

Marked and annotated region= More contextual information



Annotation Module



VLM

How?

Feeding this into the VLM → Leads to more focused and accurate answers.

Basics

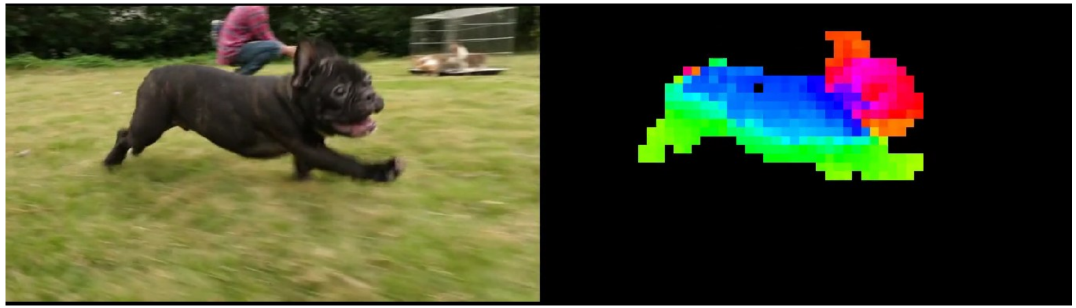
Method Selection — Object Recognition Annotation

This is exactly what we want!

Method	Method Categories	Methodological overview	Pros	Cons	Result	Segmentation	Recognition
OS2D – One-Stage One-Shot Object Detection by Matching Anchor Features(2020)	Deep Learning	Combines dense anchors with semantic transforms for category-free detection; localization + recognition via feature-level matching.	Joint detect and recognize; no retrain; general, efficient.	Lower accuracy than two-stage; scale-sensitive; needs extra processing.	Not suitable for the long distance object and objects with fewer features.	✗	✅
YOLOv11(2024)	Deep Learning	A faster, more accurate upgrade of YOLOv8, using lighter blocks, spatial attention, and better multi-scale pooling to detect objects.	Joint detection and classification and segmentation and pose estimation; high speed and accuracy.	Sees 80 classes; unseen misidentified; needs extra data.	Unfined-tuned YOLO suits daily inspection better than mechanical engineering.	✅	✅
DINO-X(detect with text prompt)	Deep Learning	DINO-X: multi-prompt, Grounding-100M trained, unifies detection & segmentation; **Pro** for accuracy, **Edge** for speed	flexible prompts, strong long-tail detection, unified multi-tasks, Pro = accurate, Edge = fast.	heavy, segmentation weaker than SAM, training costly	Unsuitable.	✅	✅
DINOV2	Deep Learning	A self-supervised ViT trained on 142M curated images with efficient scaling tricks and distillation, yielding robust frozen features for many vision tasks	Strong generalization. Robust features. Lower compute cost.	Extremely high training cost. Slightly weaker than task-specific models.	Suitable ✓	✅	✅
...	...	...	...	...	...	...	...

DinoV2 — Effect

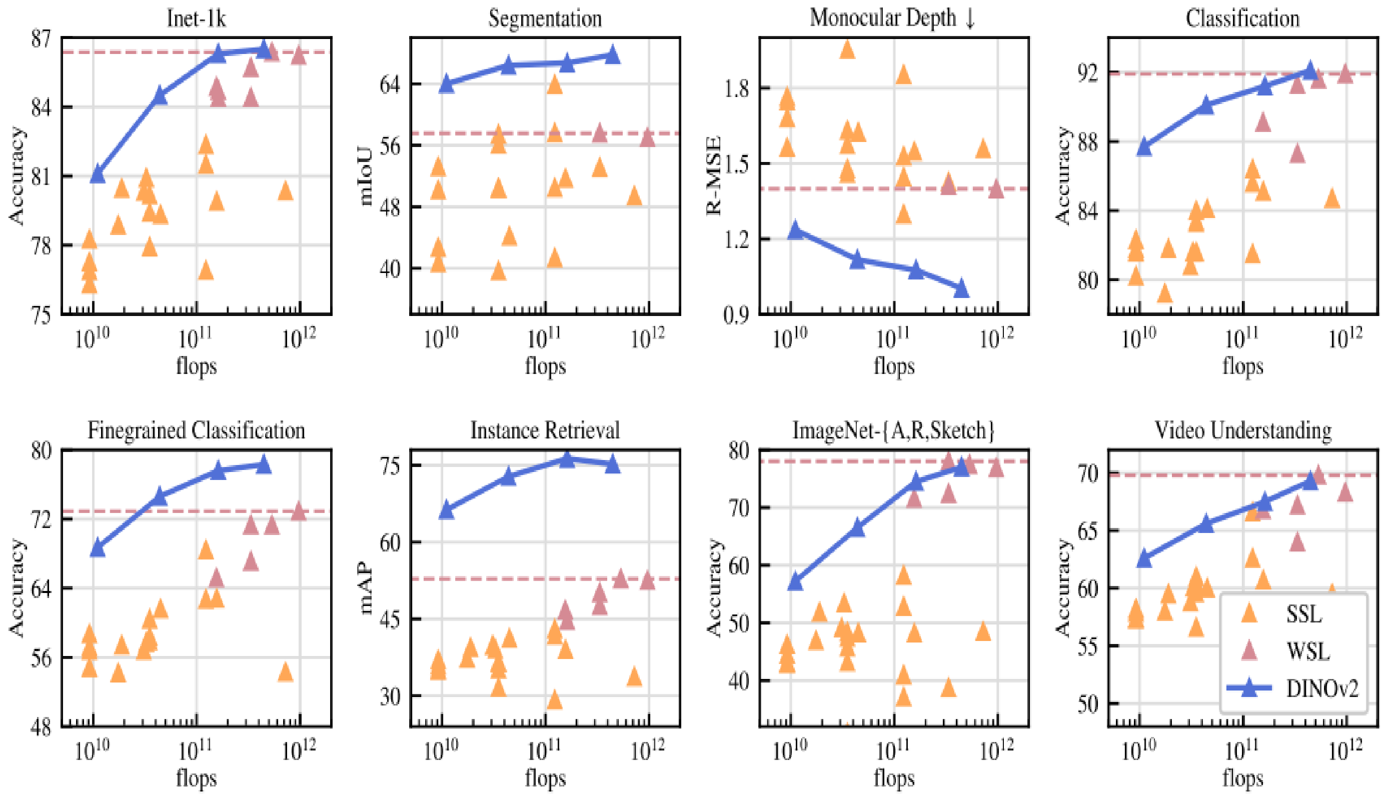
Segmentation:



Recognition:

→ **Object Type:**
Staffordshire bullterrier

DinoV2 — Benchmark

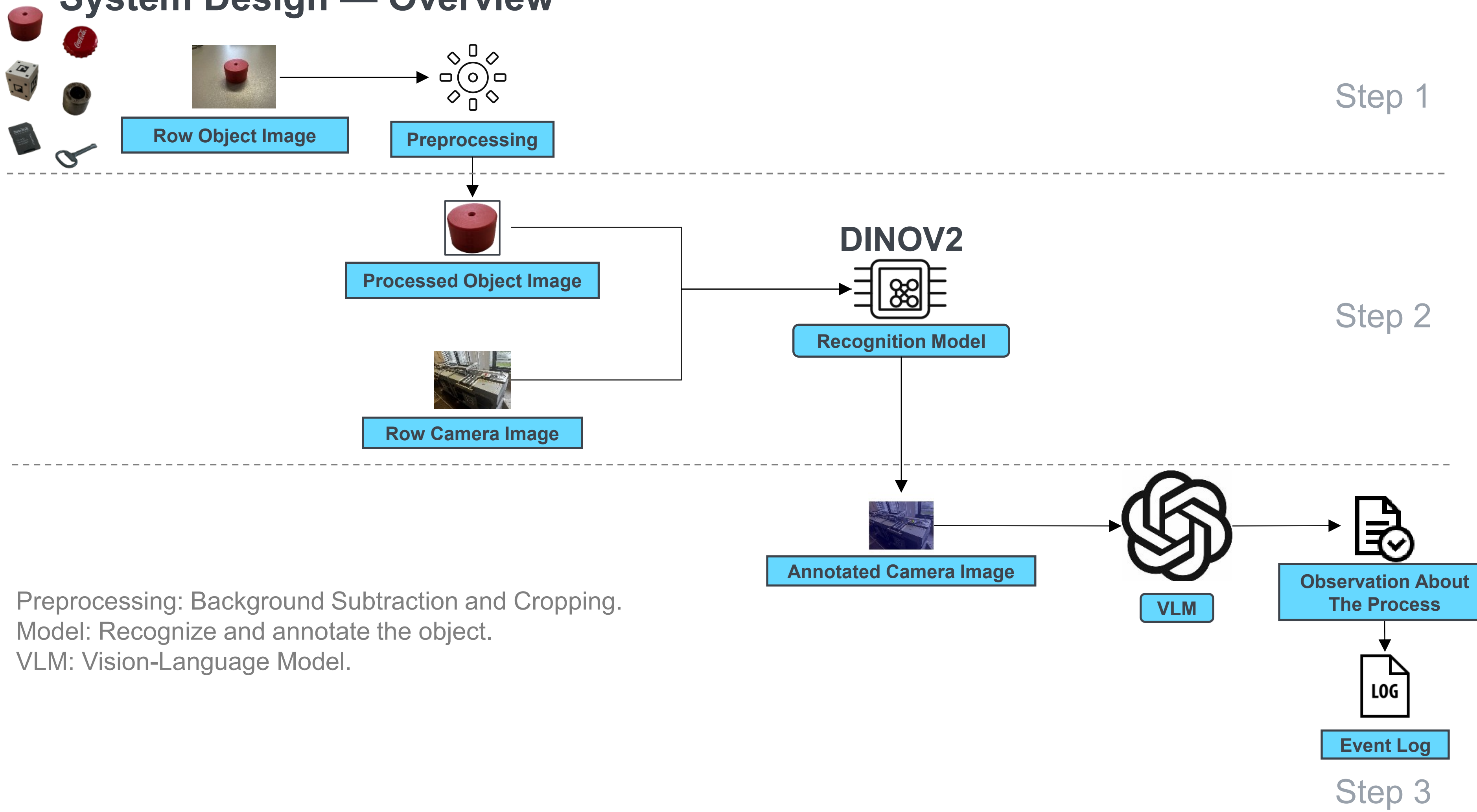


M. Oquab et al., “DINOv2: Learning Robust Visual Features without Supervision,” arXiv preprint, https://arxiv.org/abs/2304.07193, 2023.

These highlight the powerful feature extraction capabilities of DINOv2.

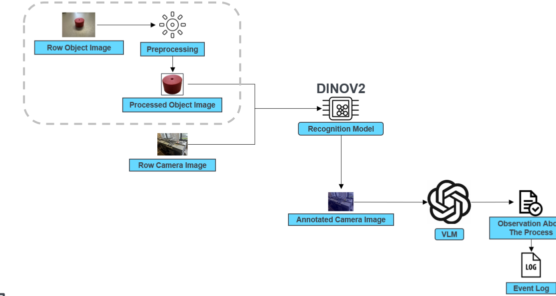
System Design

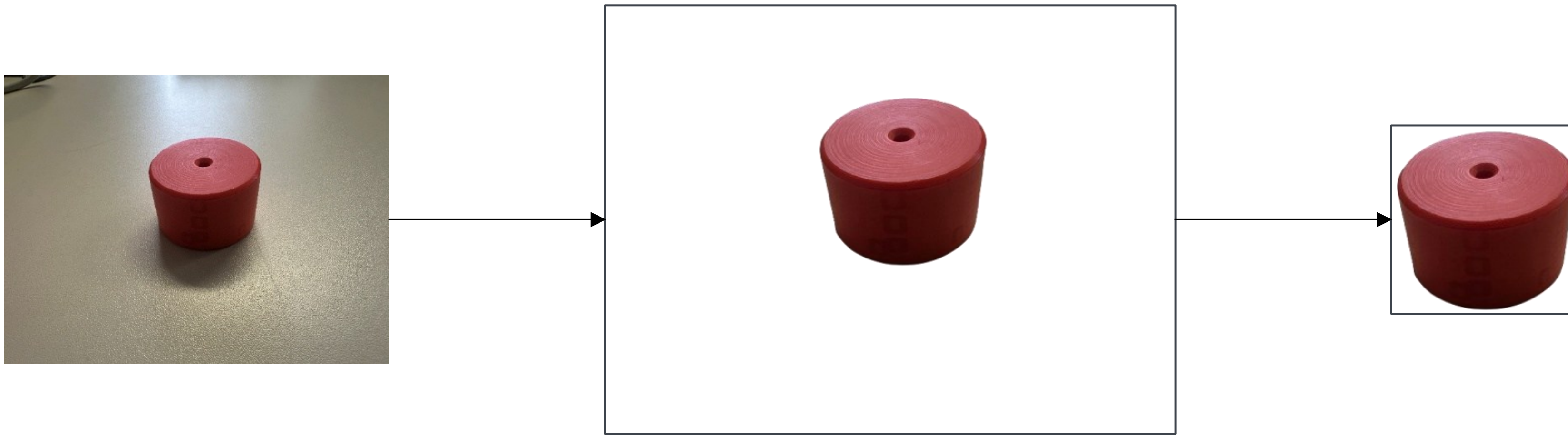
System Design — Overview



Preprocessing: Background Subtraction and Cropping.
Model: Recognize and annotate the object.
VLM: Vision-Language Model.

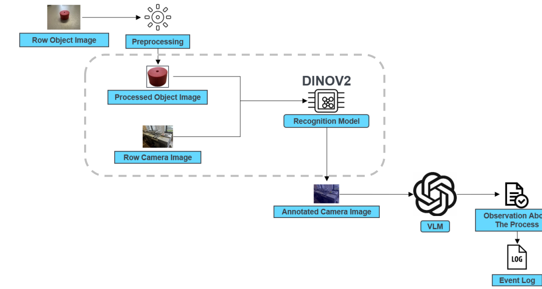
Step 1 — Preprocessing

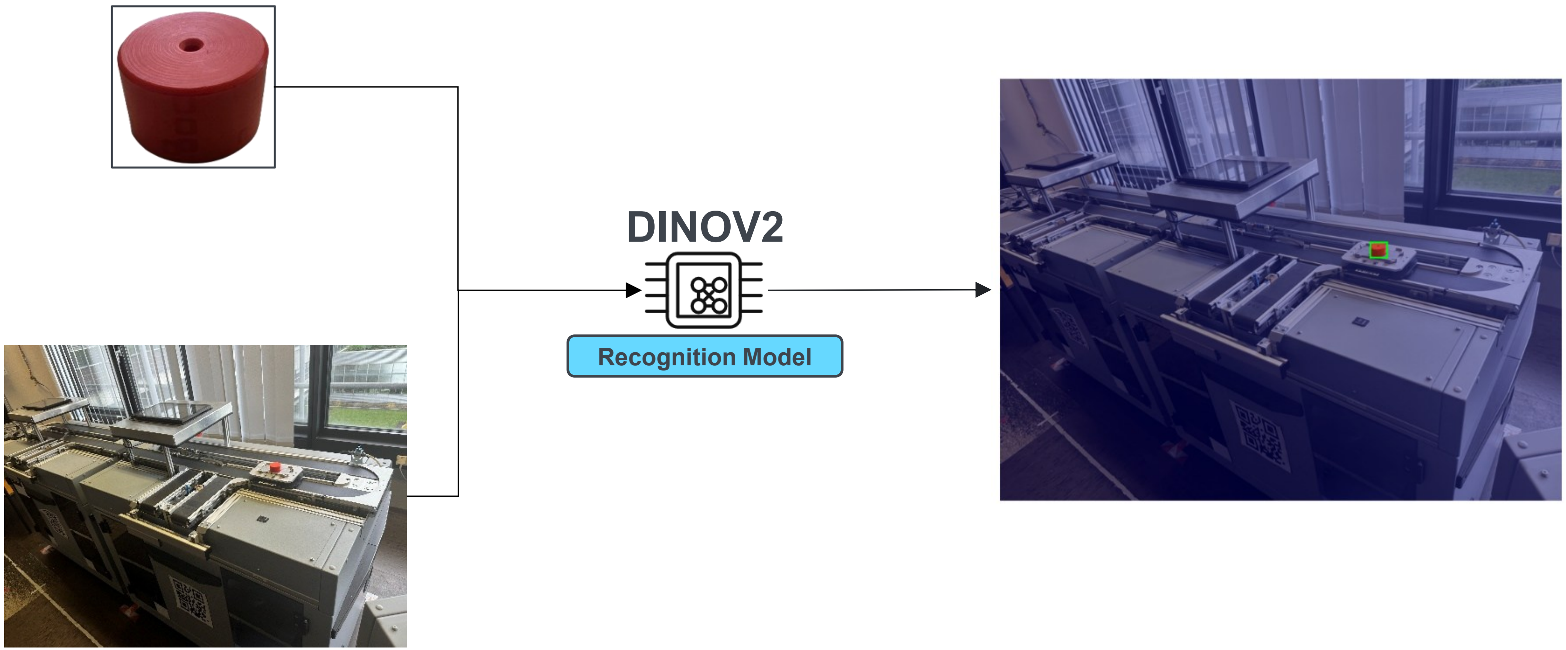




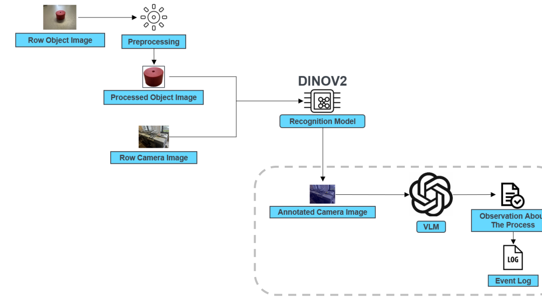
Background removal: Retaining background introduces irrelevant semantics, causing embeddings to be biased toward the scene rather than the object. Removing it yields cleaner representations that emphasize object semantics.
Object filling the template: If the object occupies only a small fraction of the template, its semantics may be overwhelmed by background features. Ensuring the object fills the template increases its contribution, enforces scale consistency, and enhances discriminative power.

Step 2 — Object Recognition Annotation





Step 3 — Reasoning



VLM

Input
One production-scene image (optionally with a bounding box around the target workpiece).

Requirements
1) Workpiece State
- On the conveyor belt
- (Optionally) on its designated carrier
- In standby, awaiting processing

2) Conveyor Position & Movement (counterclockwise system)
- Has just passed right turning part
- Continues leftward along the belt
- Is approaching the next station

3) Process Objectives (reference only; not for scoring)
- Positional accuracy of the workpiece
- Continuity of material flow

4) If "Fail", give one brief, position-related reason.
- If a C2/C3 position comparison is possible, explicitly mention it (e.g., "workpiece still on C2, not on C3").
- If not, provide a simple position-related reason (e.g., "not on carrier", "before station", "moving leftward").

Evaluation Steps
1) Verify Workpiece State (belt / carrier / standby).
2) Verify Position & Movement relative to the station and rightward motion toward the right turning part (under counterclockwise flow).
3) Decision Rule: Output "Success" only if all required items in (1) and (2) are satisfied from the image; otherwise output "Fail".
4) If "Fail", give one brief, position-related reason.
- If a C2/C3 position comparison is possible, explicitly mention it (e.g., "workpiece still on C2, not on C3").
- If not, provide a simple position-related reason (e.g., "not on carrier", "before station", "moving leftward").

Answer Format (strict JSON)
Return only one JSON object in the exact format specified (no extra text, no explanation):
```
{
 "Result of planned process": "Success or Fail",
 "Description": "One reason in a properly capitalized English sentence (position-related)"
}
```

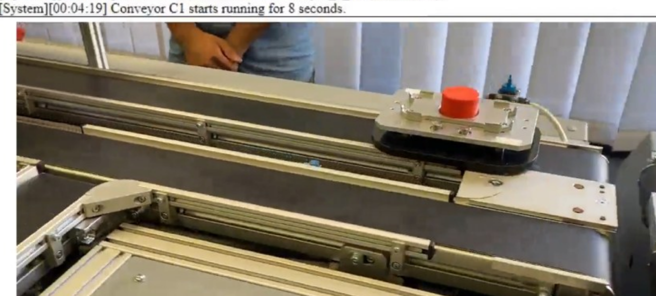
```
{
  "Result of planned process": "Fail",
  "Description": "workpiece still on C2, not on C3"
}
```

```
[00:00:35] Fail. workpiece still on C2, not on C3
```

Adding into raw event log.

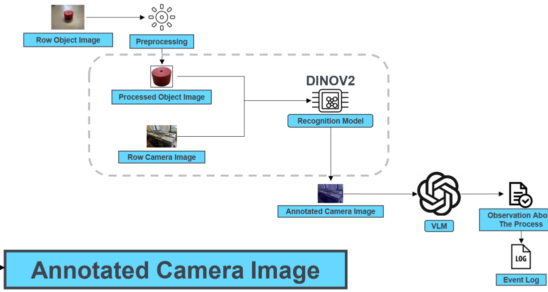
Event log memory

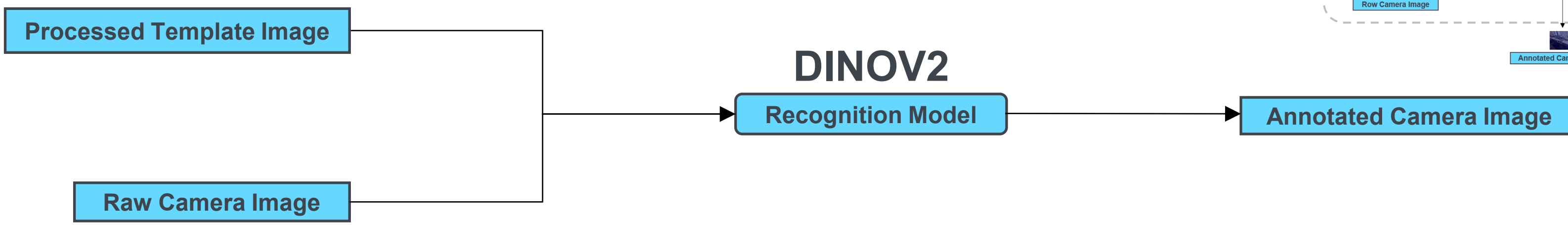
[Painting Station][System][00:04:18] Holder H2 is raised.
[Painting Station][System][00:04:18] BG57 detects a workpiece at the outlet of conveyor C1.
[Painting Station][Operator][00:04:18] Task completion for Painting Station: load and coat the workpiece "white cylinder" with red paint.
[MES][System][00:04:18] Task completion for Painting Station: load and coat the workpiece "white cylinder" with red paint.
[MES][Manager][00:04:19] Task assigned to CNC Station: drill the workpiece "red cylinder".
[CNC Station][System][00:04:19] BG56 detects a workpiece at the infeed of conveyor C1.
[CNC Station][Operator][00:04:19] CNC Station calls function: conveyor_1_run('forward', 8).
[CNC Station][System][00:04:19] Conveyor C1 starts running for 8 seconds.



Experiments

Evaluation





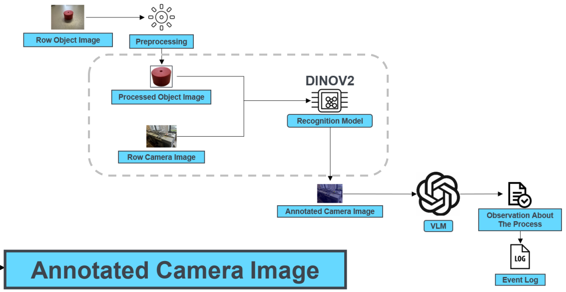
Success:

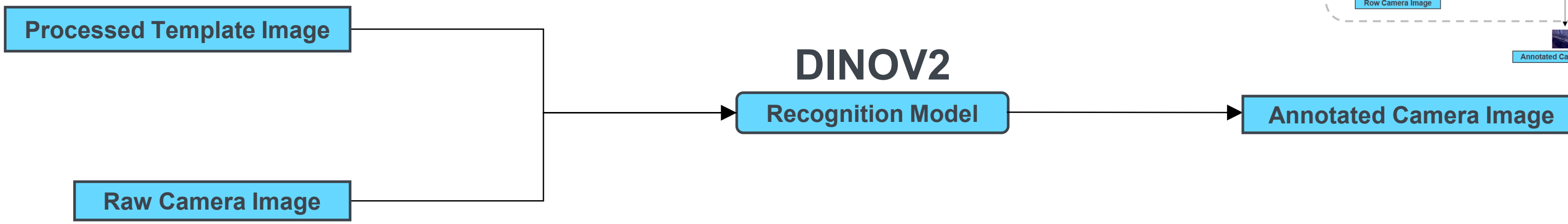
Fail:

Number of Successes: 23 **Success Rate:** 95.8%

Number of Fails: 1 **Fail Rate:** 4.2%

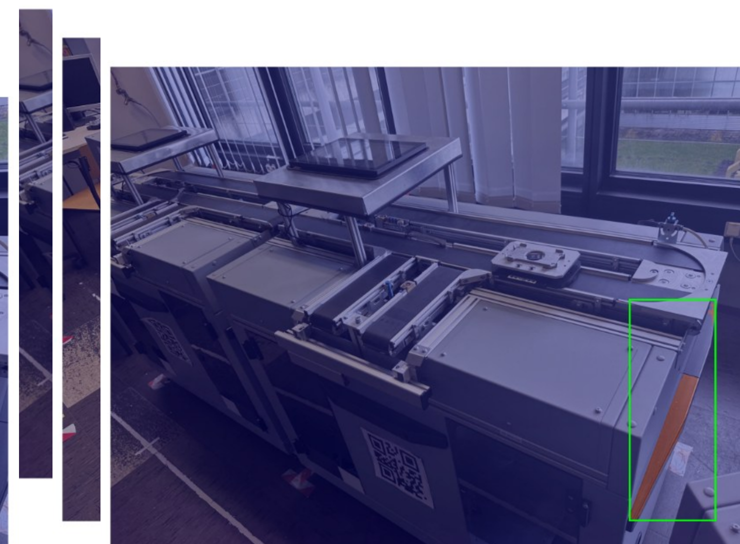
Evaluation — Long Distance Range





Success:

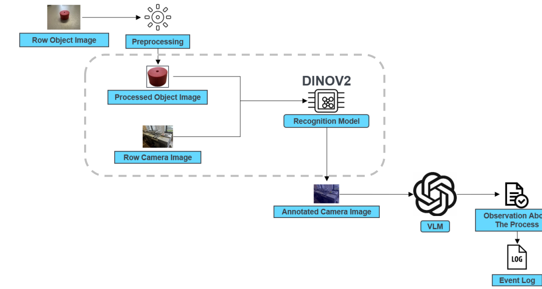
Fail:



Number of Successes: 14 **Success Rate:** 77.8%

Number of Fails: 4 **Fail Rate:** 22.2%

Evaluation — Summary



The evaluation set consists of 42 images containing six random objects, with 24 taken at close range and 18 at long range.

Recognition of different objects:

100%

100%

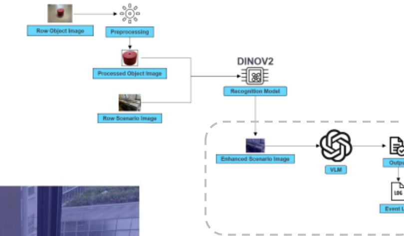
60%

100%

100%

83%

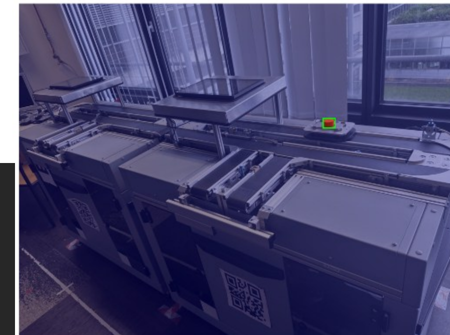
Outcome\Range	close range	long range
Correct	23	14
Wrong	1	4
Success Rate	95.8%	77.8%



Comparison — VLM Outputs (Example)

Without Annotation:

With Annotation:



Input
One production-scene image (optionally with a bounding box around the target workpiece).

Requirements
1) Workpiece State
- On the conveyor belt
- (Optionally) on its designated carrier
- In standby, awaiting processing

2) Conveyor Position & Movement (counterclockwise system)
- Has just passed right turning part
- Continues leftward along the belt
- Is approaching the next station

3) Process Objectives (reference only; not for scoring)
- Positional accuracy of the workpiece
- Continuity of material flow

Evaluation Steps
1) Verify Workpiece State (belt / carrier / standby).
2) Verify Position & Movement relative to the station and rightward motion toward the right turning part (under counterclockwise flow).
3) Decision Rule: Output "Success" only if all required items in (1) and (2) are satisfied from the image; otherwise output "Fail".
4) If "Fail", give one brief, position-related reason.
- If a C2/C3 position comparison is possible, explicitly mention it (e.g., "workpiece still on C2, not on C3").
- If not, provide a simple position-related reason (e.g., "not on carrier", "before station", "moving leftward").

Answer Format (strict JSON)
Return only one JSON object in the exact format specified (no extra text, no explanation):
```
{
 "Result of planned process": "Success or Fail",
 "Description": "One reason in a properly capitalized English sentence (position-related)"
}
```

```json
{
    "Result of planned process": "Fail",
    "Description": "No workpiece is detected on the conveyor belt."
}
```

```
[00:00:40] Fail. No workpiece is detected on the conveyor belt.
```

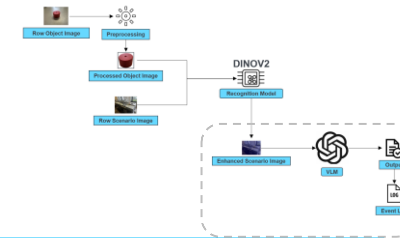
Wrong!

```json
{
    "Result of planned process": "Success",
    "Description": "Workpiece is on its carrier on C3 just past the right turning section, positioned leftward and approaching the next station."
}
```

```
[00:00:40] Success. Workpiece is on its carrier on C3 just past the right turning section, positioned leftward and approaching the next station.
```

Correct!

Comparison — VLM Outputs (Summary)



Close range

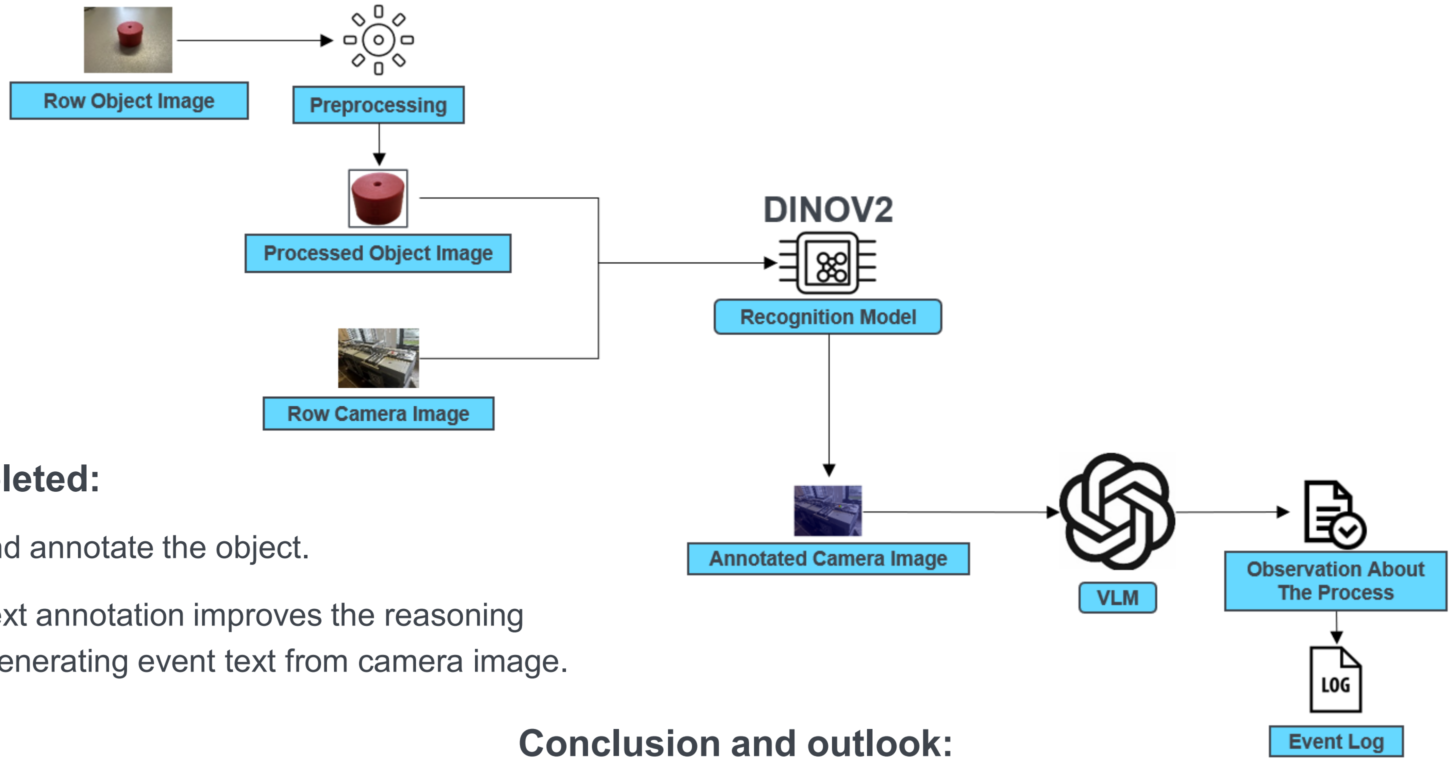
Outcome\Stages	Without Annotation	With Annotation	Event log
Correct	16	19	[00:00:29] Success. Workpiece is on its carrier…
Wrong	8	5	[00:00:30] Fail. no workpiece visible on the conveyor.
Success Rate	66.7%	79.2% (+12.5%)	/

Long range

Outcome\Stages	Without Annotation	With Annotation	Event log
Correct	1	14	[00:00:40] Success. Workpiece is on its carrier on C3…
Wrong	17	4	[00:00:30] Fail. no workpiece visible on the conveyor.
Success Rate	5.6%	77.8% (+72.2%)	/

Conclusion & Future Work

Conclusion & Future Work



Tasks Completed:

- Recognize and annotate the object.
- proven: context annotation improves the reasoning accuracy in generating event text from camera image.

Limitation:

- 78.6% accuracy in event text generation is still not reliable enough

Conclusion and outlook:

- The result improvement is significant (12 – 72 % increase in accuracy)
- Further improvement is still need for real application beyond the state-of-the-art methods

University of Stuttgart
Institut of Industrial Automation
and Software Engineering

Thank you!

Zhongxin Cao

e-mail st180604@stud.uni-stuttgart.de
phone +49 (0) 711 685-
fax +49 (0) 711 685-

University of Stuttgart
Institute of Automation and Software Systems
Pfaffenwaldring 47, 70550 Stuttgart