International Workshop by WINS
Introduction to the Role of AI in Strengthening the Security of Nuclear Facilities

6-8 February 2024, Vienna, Austria

Proposal and Development of Methods for Detection of Malicious Acts

Kazuyuki Demachi
Associate professor,
Department of Nuclear Engineering and Management
School of Engineering, The University of Tokyo

東京大学 THE UNIVERSITY OF TOKYO

nem 原子力国際専攻 Nuclear Engineering and Management

https://www.demachilab.org/

Contents

Biography of Kazuyuki Demachi

History

1997 March: Graduate the doctor course of Department of System Innovation, School of Engineering, the University of Tokyo (Doctor Degree of Engineering)
1997 March-1998 April: Assistant of Nuclear Engineering Research Laboratory, School of Engineering, the University of Tokyo
1998 May-2000 March: Lecturer of Nuclear Engineering Research Laboratory, School of Engineering, the University of Tokyo
2000 April-2012 March: Associate professor of Nuclear Professional School, School of Engineering, the University of Tokyo
2012 April-present: Associate professor of Department of Nuclear Engineering and Management, School of Engineering, the University of Tokyo

Recent Research Topics

- **Nuclear Security (Detection of insider, Cyber-attack, BDBT countermeasure)**
- Nuclear Power Plant Maintenance Technology **by AI**.
- Medical Imaging Technology of Lung Tumor **by AI**

1. Introduction

(1) Ever-changing nuclear security threats

Nuclear security threats are ongoing and continue to change even nowadays

~INFCIRC/225/Rev4

Nuclear terrorism
① Theft of nuclear weapons
② Theft of nuclear material for nuclear weapons
(③Sabotage to NPP & transport)

After 911 US attack in 2001

diversifed

INFCIRC/225/Rev5

In 2011

Terrorism against society
① Theft of nuclear weapons
② Theft of nuclear material for nuclear weapons
③ **Theft of RI for dirty bomb**
④ **Sabotage to NPP & transport**

Furthermore, in recent years

more sophisticated

New Threats means
① Insider
② Cyber attack
③ Stand-off attack

In 2022,

expanded

INFCIRC/225/Rev6 ?

Paradigm Shift occurred
① Wartime nuclear security
② Nuclear security for BDBT*

The Zaporizhia nuclear power plant in Ukraine was occupied by Russian forces

(2) Why "detection" should be focused on ?

Four stages of Physical Protection (PP):

Deterrence → Detection → Delay → Response

make the enemy give up

malicious behavior detection

earning time

Neutralization by force (police, military)

*1: Incident and Trafficking Database by IAEA
*2: Central Alarm Station

- In these, detection is the **bottle-neck** of PP because delay and response do not work if detection fails.
- **Enhanced detection** is critical for physical protection against 'new threats' to work
- Then, how can we detect the "new" threats?

(1) Against Insider

- ➢ Insider has free access, so deterrence is invalid.
- ➢ ITDB report*1: About 100 incidents/year, most of them are by insiders
- ➢ Now, **huge numbers** of surveillance camera are monitored and observed by human eye in CAS*2
- ➢ It is classical and vulnerable.
- ➢ New technology to help detecting malicious behavior **as a primely screening** is necessary.
- ➢ Images are the form that contains the most information about "human behavior", ⇒ **detection using "image" was developed.**

(2) Against Cyber attack

- ➢ Cyber space : network → control → physical layers
- ➢ Current: Monitoring mainly NW traffic and server activity on NW layer.
- ➢ The order of investigation: equipment → control → cyber. Needs a half day.
- ➢ Early detection targeting cyber attack on the physical layer is necessary.

(3) Against Stand-off

- ➢ Long-range attacks outside monitoring areas
- ➢ One of the effective means = earning time until Force arrives
- ➢ Early detection is the key to earning time
- ➢ Detection of "human behavior" ⇒ "image"

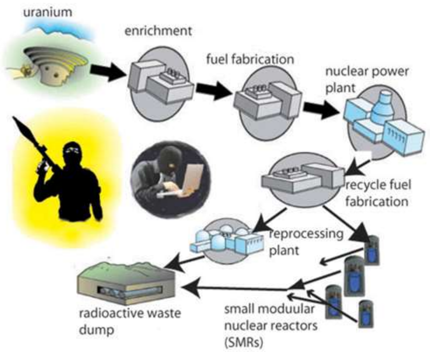
(3) Detection of Malicious Action

(1) External intrusion: Breaching or infiltrating into the physical boundaries **Common** **Detect-in-advance**

- Fence climbing

- Fence destroying

- Weapon holding



- …

(2) Insider threat: Attending inside the nuclear power plants. **Complex** **Serious**

- Theft

- Violence

- Weapon holding
- …

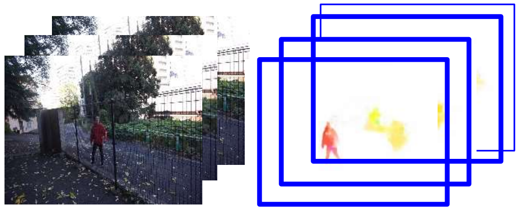
2. Methods for detection of malicious actions

could be divided into two : **video-centered** and **human-centered**

(1) Video-centered method

Black-box

Easy-to-implement



Raw / optical flow frame sequences

Abnormal detection model → Reasoning result

→ Using one black-box model to directly get the reasoning result of the whole video

(2) Human-centered method

Explainable

Complex

Raw frame sequences

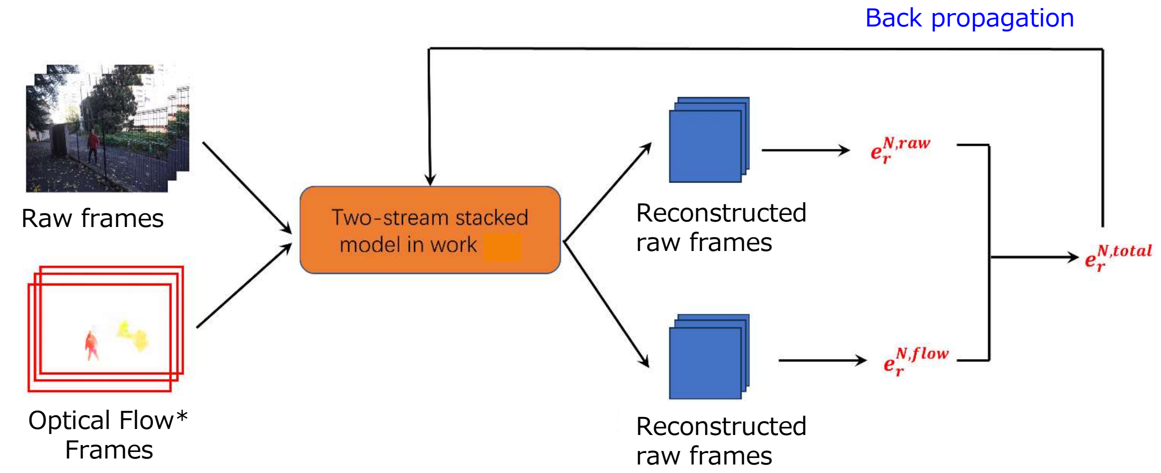
Computer vision module (CVM) → Reasoning module (RM) → Reasoning result

Using CV-technologies to extract **multi-type human actions**

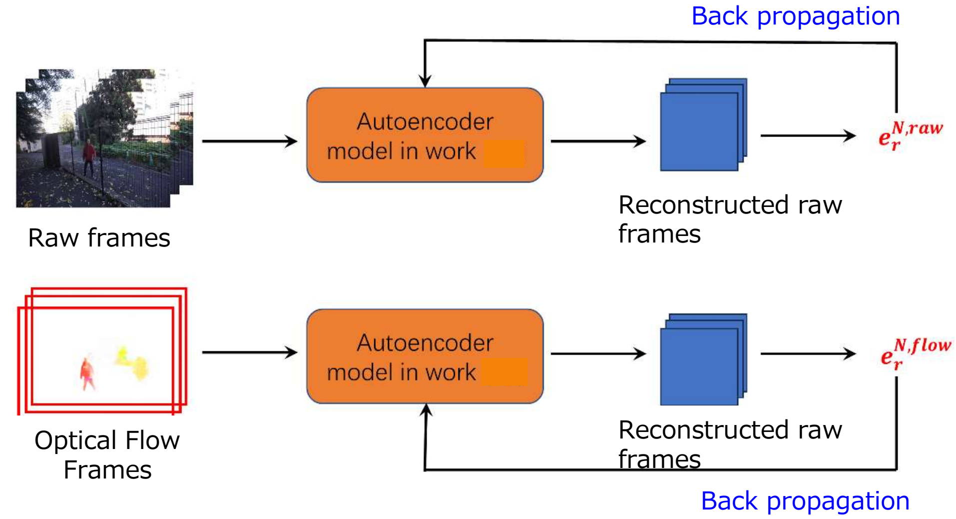
Development of specific reasoning methods

(1) Video-centered method: Label-specific method

1) Single-stream: A two-stream stacked model is trained for each label-specific branch



2) Double-stream: Two sub-autoencoder models are separately trained



Deep learning models were trained to output high e_r (=label) for Raw frames and optical flows with abnormal behavior.

*Optical Flow：Movement vector of the pixel that changed

(1) Video-centered method: Label-specific method

Experiment & results

Scenario 1:Fence climbing

Scenario 2:Wire net cutting

Scenario 3:Weapon holding

Scenario 4:Armed boundary sabotage

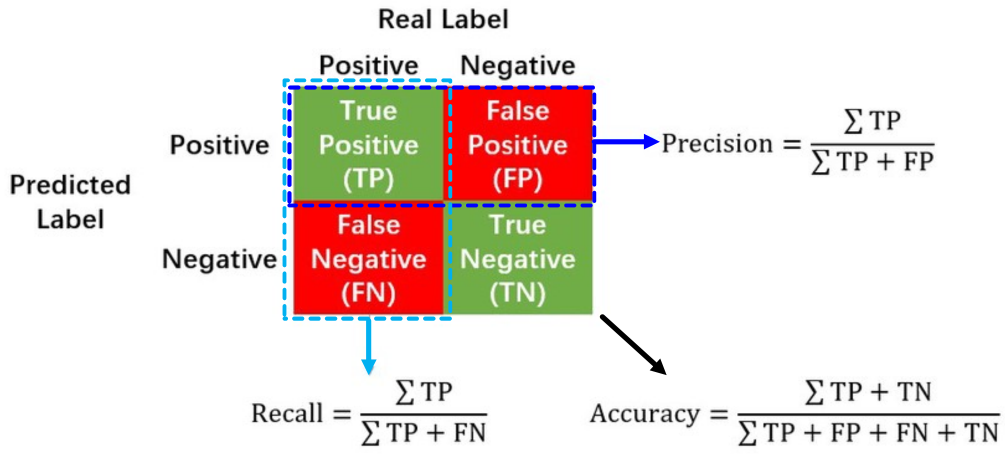
Dataset	Video number	Time length
Training & Validation	90	32 min 31 s
Testing	55	14 min 49 s

Self-collected dataset: 4 abnormal scenarios + normal status

(1) Video-centered method: Label-specific method

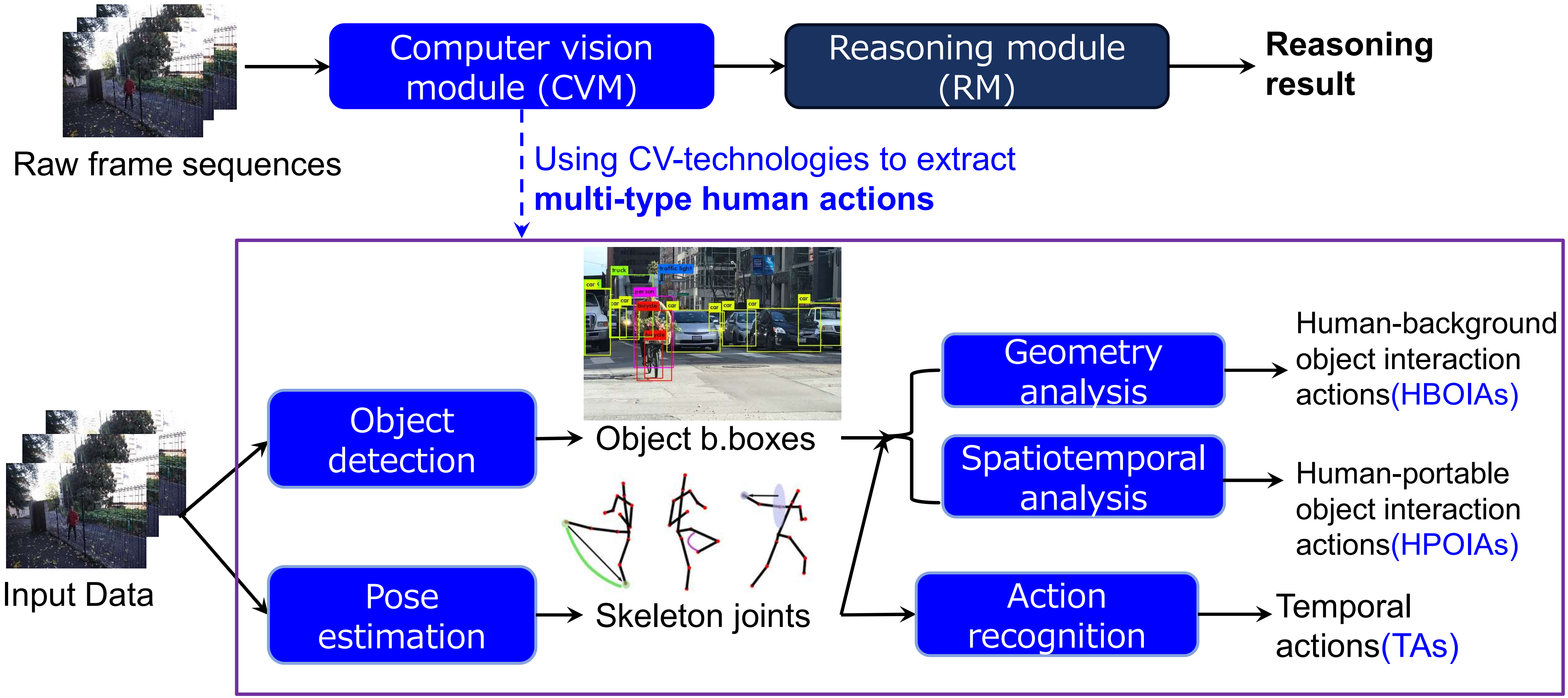
Experiment & results

Method	Threshold	Precision	Recall	Accuracy	Supplement
Traditional	1e-1	0	0	0.5273	All normal
	1e-2	0	0	0.5273	All normal
	1e-3	0	0	0.5273	All normal
	1e-4	0.4727	1	0.4727	All abnormal
	1e-5	0.4727	1	0.4727	All abnormal
	1e-6	0.4727	1	0.4727	All abnormal
One-stream (Proposed)		**0.7222**	**1**	**0.8182**	—
Two-stream (Proposed)		**0.7500**	**0.9231**	**0.8182**	—



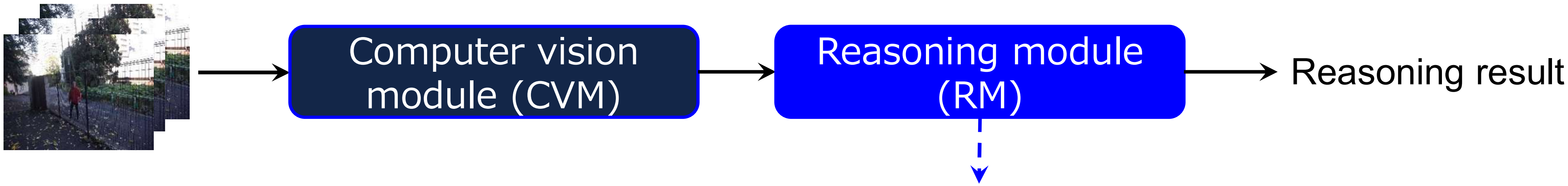
Our proposed method could solve the problems better than the traditional models.

(2) Human-centered method



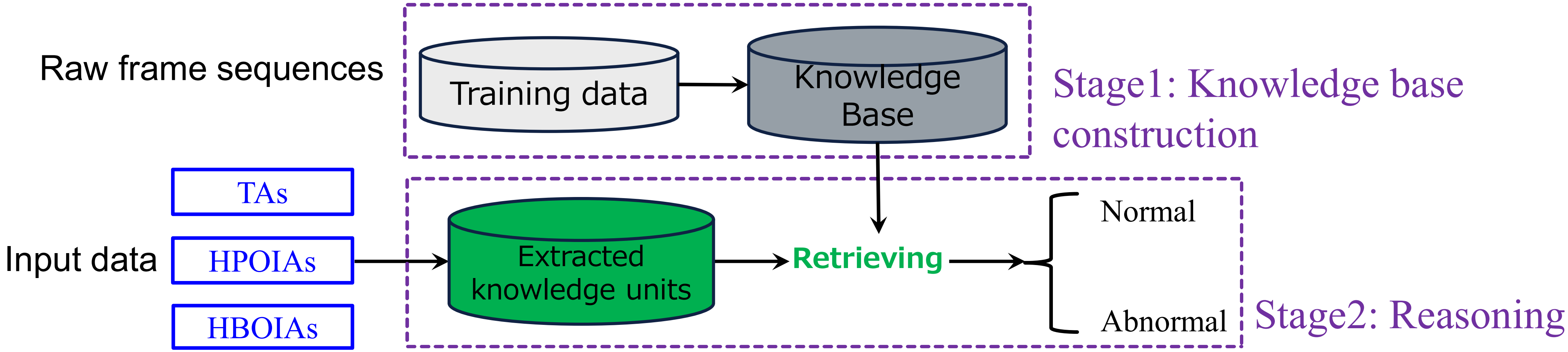
CVM includes 5 sub-calculation processes, while the output contains three types of human actions.
However, the development of novel explainable reasoning methods for RM is more significant.

(2-1) Data-based reasoning method



Development of specific reasoning methods

Essence: Statistical models



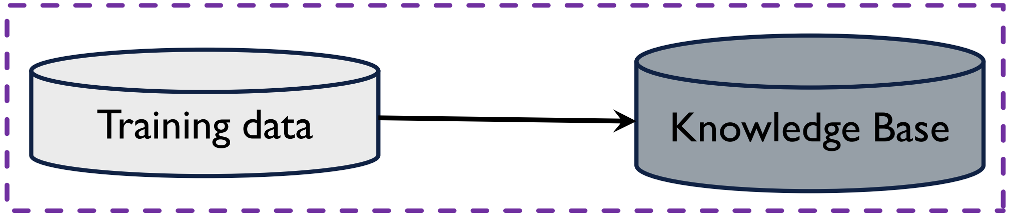
Stage1: Knowledge base construction

→ Aiming to obtain the knowledge base comprised by malicious features

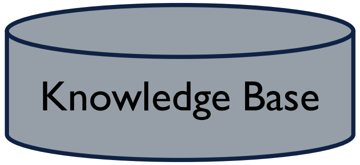
Stage2: Reasoning

→ Retrieving the existence of extracted knowledge units on knowledge base

(2-1) Data-based reasoning method



Stage1: Knowledge base construction



- Self-malicious features: $unit_a$
- Co-occurrence malicious features: $unit_a + unit_b$
  - → detected in the same frame
- Sequential malicious features: $unit_a \rightarrow unit_b$
  - → detected in sequential frames

$unit$: TAs, HPOIAs, HBOIAs

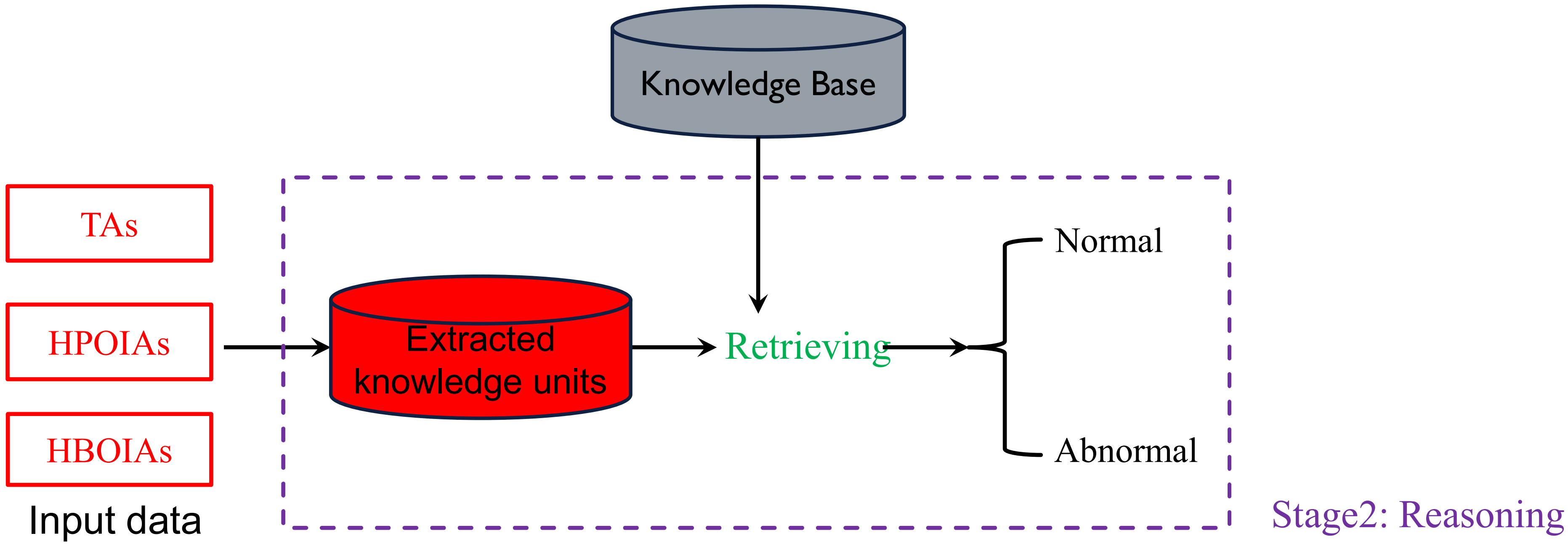
(1) For all the candidate malicious features, they should be judged:

$$\frac{n_{self}^{malicious}}{n_{self}^{total}} = 1 \qquad \frac{n_{co}^{malicious}}{n_{co}^{total}} \geq \alpha_1 \qquad \frac{n_{sequential}^{malicious}}{n_{sequential}^{total}} \geq \alpha_2$$

(2) Besides, there is also filtering process:

For $feature_{co} = unit_a + unit_b$, if $unit_a / unit_b$ belongs to self-malicious feature, then delete $feature_{co}$
Similar for $feature_{sequential}$

(2-1) Data-based reasoning method



After knowledge base is constructed, the reasoning stage is relatively easy

(1) Any element in extracted knowledge units exists in knowledge base → **Abnormal**

(2) Otherwise → **Normal**

(2-2) Language-based reasoning method

Essence: CV + NLP, imitating the function of language center

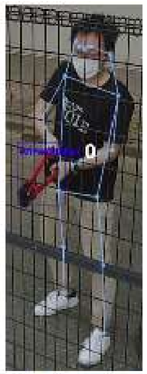
Output of CVM → Natural language generation → Intent-appended natural languages → Reasoning results

Pipeline 1: Intention-induced natural language generation (INLG)

Output of CVM → Natural language generation → Intent-irrelevant natural languages → Natural language inference → Reasoning results

Pipeline 2: Joint natural language generation & inference (JNLGI)

Difference
- Generated languages
- Natural language inference



Scenario

（1） Intent-appended natural language：Human is walking close to the wirenet with a wirecutter in hand, which is related to external intrusion.

（2） Intent-irrelevant natural language：Human is walking close to the wirenet with a wirecutter in hand.

(2-2) Language-based reasoning method

Model1: Natural language generation

Table 1. An E2E data instance. The meaning representation appears in the dataset once for each reference sentence.

Meaning Representation	References
name[The Wrestlers], eatType[coffe shop], food[Indian] priceRange[less than L20] area[city centre] familyFriendly[yes] near[Raja Indian Cuisine]	Indian food meets coffee shop at The Wrestlers located in the city centre near Raja Indian Cuisine. This shop is family friendly and priced at less than 20 pounds. Near Raja Indian Cuisine, The Wrestlers provides the atmosphere of a coffee shop with Indian food. At less than 20 pounds, it provides a family friendly setting for its customers right in the city centre. The Wrestlers is a coffee shop providing Indian food in the less than L20 price range. It is located in the city centre. It is near Raja Indian Cuisine.

Sample

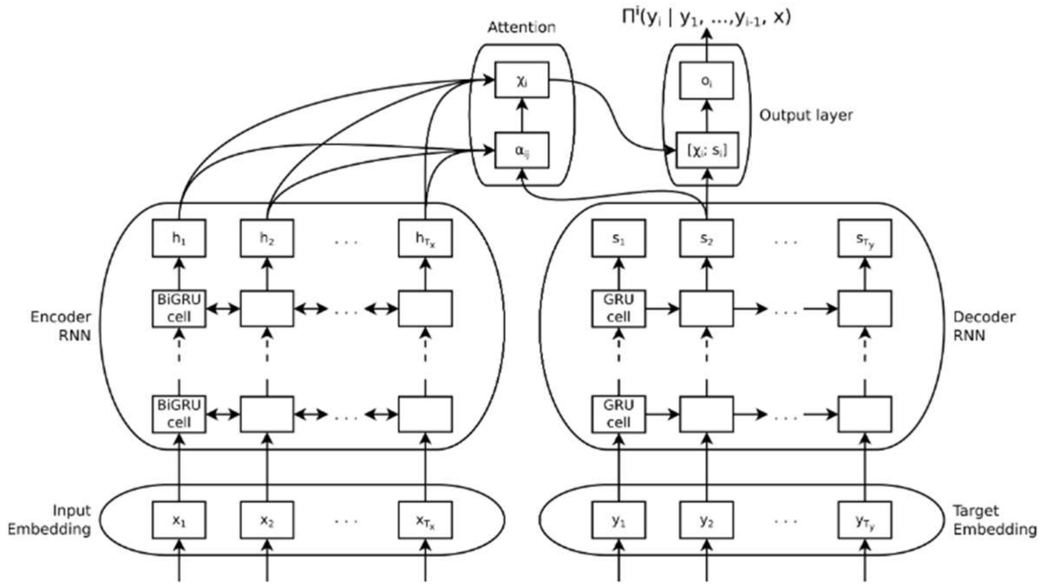


Figure 1. Encoder–decoder with attention model.

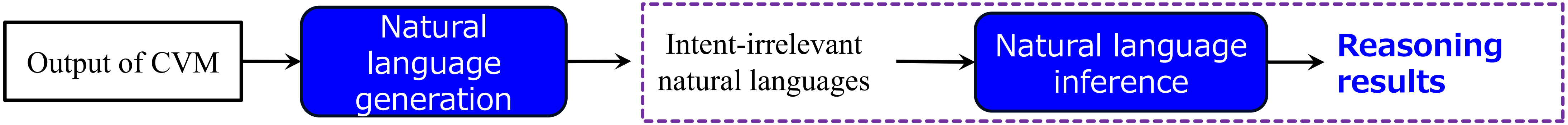
Calculation model——EDACS [1]

Input data: **Attribute[*value*] pairs**
- **TA[*TA-type*]**
- **HPOIA[*HPOIA-type*]**
- **HBOIA[*HBOIA-type*]**

Output data: **Generated languages**

[1] Bonetta, G., Roberti, M., Cancelliere, R., Gallinari, P., 2021. The rare word issue in natural language generation: a character-based solution. Informatics 8, 20.

Model2: Natural language inference



Pipeline 2: Joint Natural Language Generation & Inference (JNLGI)

Key technology: Application of ChatGPT

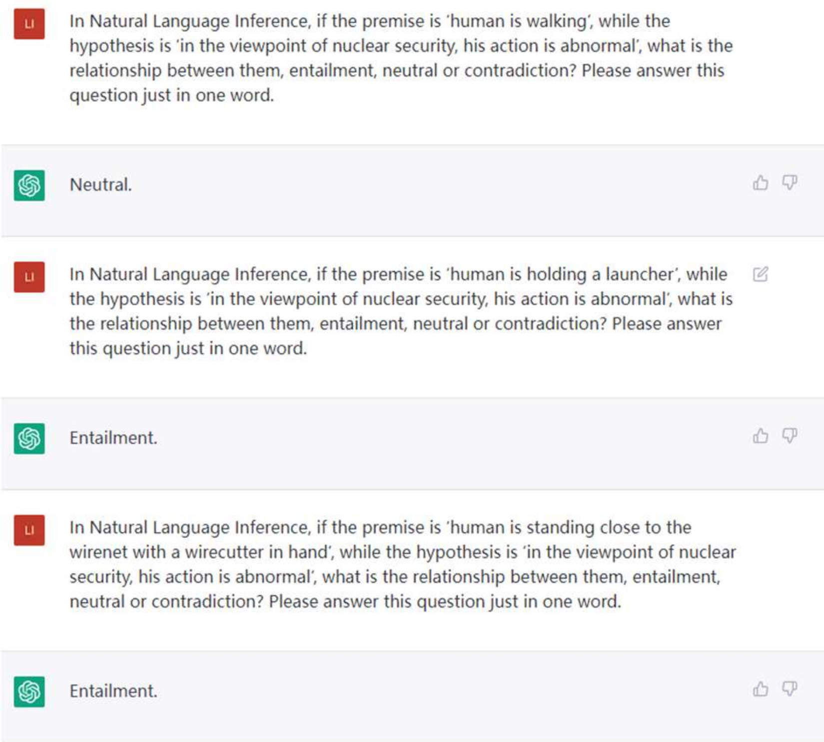
Intent-irrelevant natural languages

↓ Problem construction

*In Natural language inference, if the premise is '**generated language**', while the hypothesis is 'in the viewpoint of nuclear security, his action is abnormal', what is the relationship between them, entailment, neutral or contradiction? Please answer this question just in one word.*

↓ ChatGPT API

Answer:
- Entailment → **Abnormal**
- Neural → **Normal**
- Contradiction → **Normal**



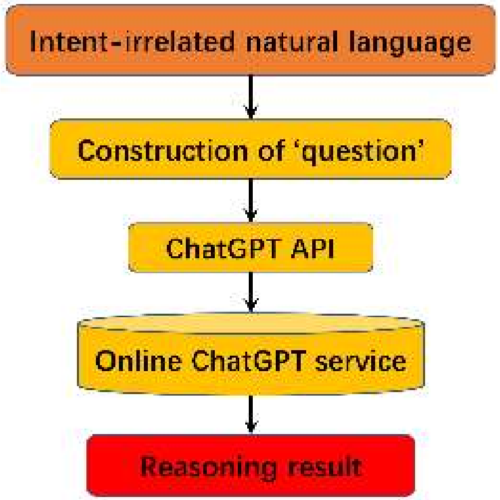
(2-2) Language-based reasoning method

Model2: Natural language inference

Pipeline 2: Joint natural language generation & inference (JNLGI)

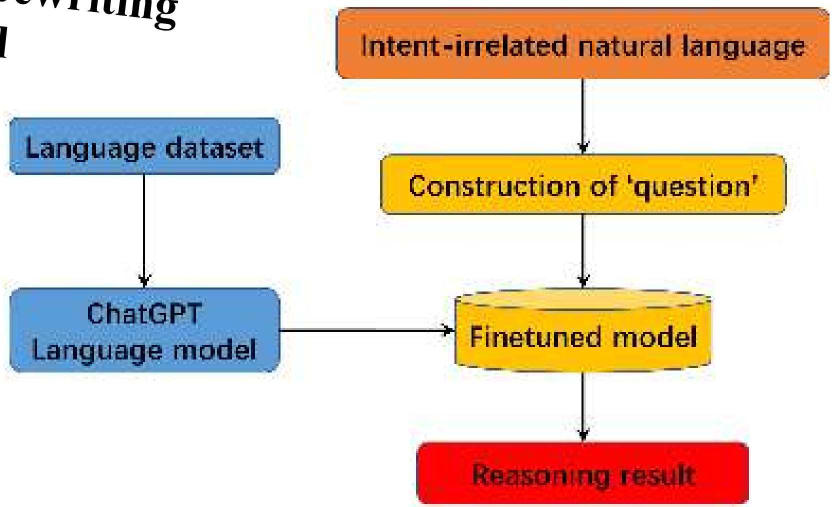
For using the ChatGPT in an automated way

Manually typewriting is unpractical



Sub-framework 1: *ChatGPT Wrapper*

→ Constructing API to original language model

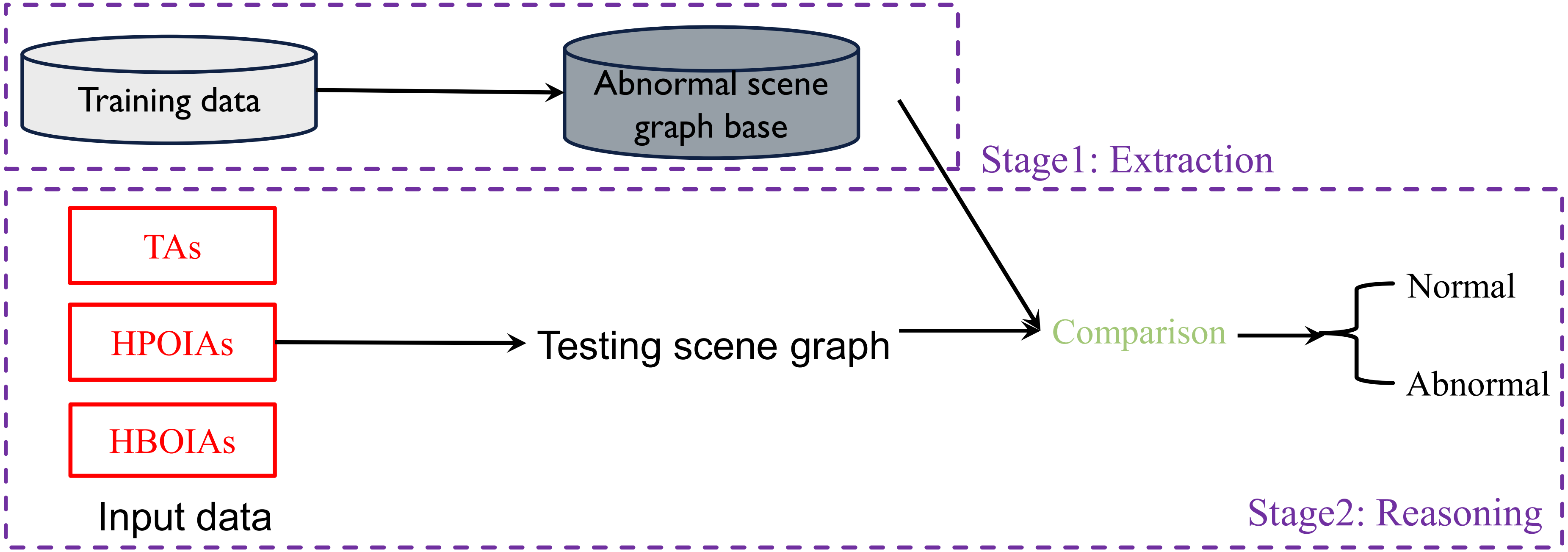


Sub-framework 2: *LMFlow*

→ Finetuning language model

(2-3) Graph-based reasoning method

Essence: Contrastive learning

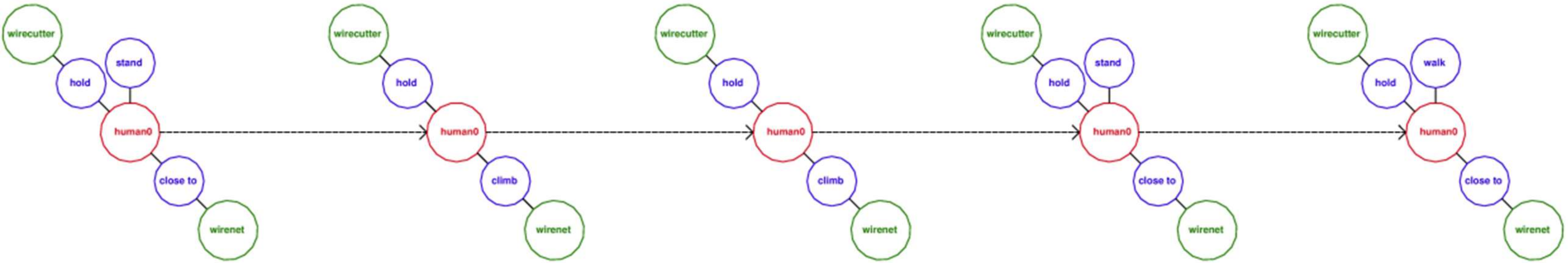


Main Points

- **(1) How to extract abnormal scene graph base?**
- **(2) How to conduct the comparison process?**

→ **Graph Similarities**

(2-3) Graph-based reasoning method



For scene graphs:

- Nodes
  - Human nodes
  - Action nodes
  - Object nodes
- Edges

Detected results ⟶ **Graph-structed data** ⟶ **Graph Similarities**

Based on **graph theory** and **discrete mathematics**

Three types of graph similarities
- (1) Jaccard coefficient
- (2) Graph edit distance
- (3) Graph Kernel value

(2-3) Graph-based reasoning method

(1) Jaccard coefficient (*JC*)

→ Ratio of common nodes

$$JC = \frac{|N_1 \cap N_2|}{|N_1 \cup N_2|} = \frac{2}{2+1+2} = 0.4$$

(2) Graph edit distance (*GED*)

→ **Minimum edit steps** to make two scene graphs congruent

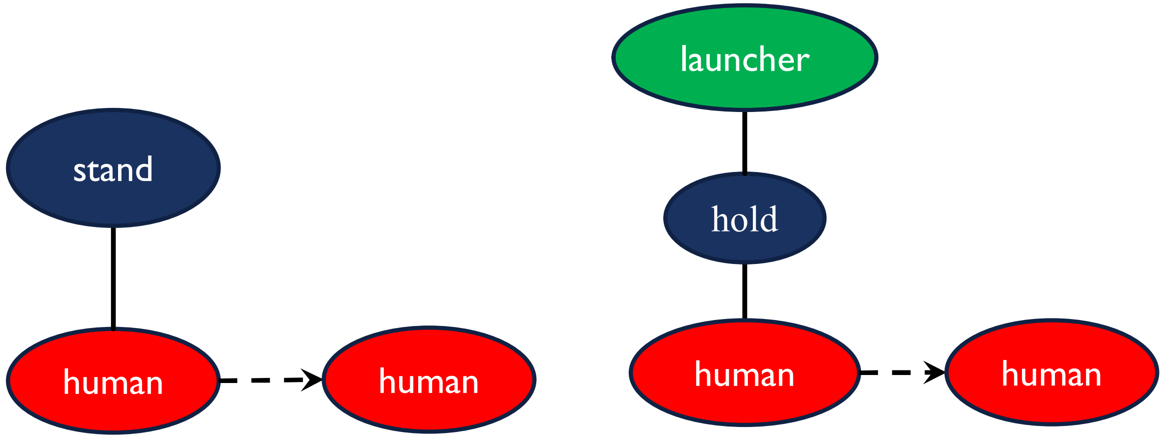
Edit step
- Replace a node
- Insert a node
- Delete a node

Finally, normalization would be used

$$RGED = \frac{GED \times 2}{|N_1| + |N_2|} = \frac{2 \times 2}{3+4} = \frac{4}{7}$$

(3) Graph Kernel value

→ Inner product of projection representations in Hilbert space

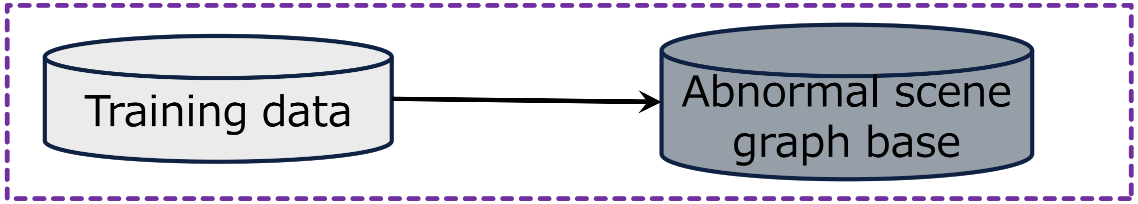


Scene graphs #1

Scene graphs #2

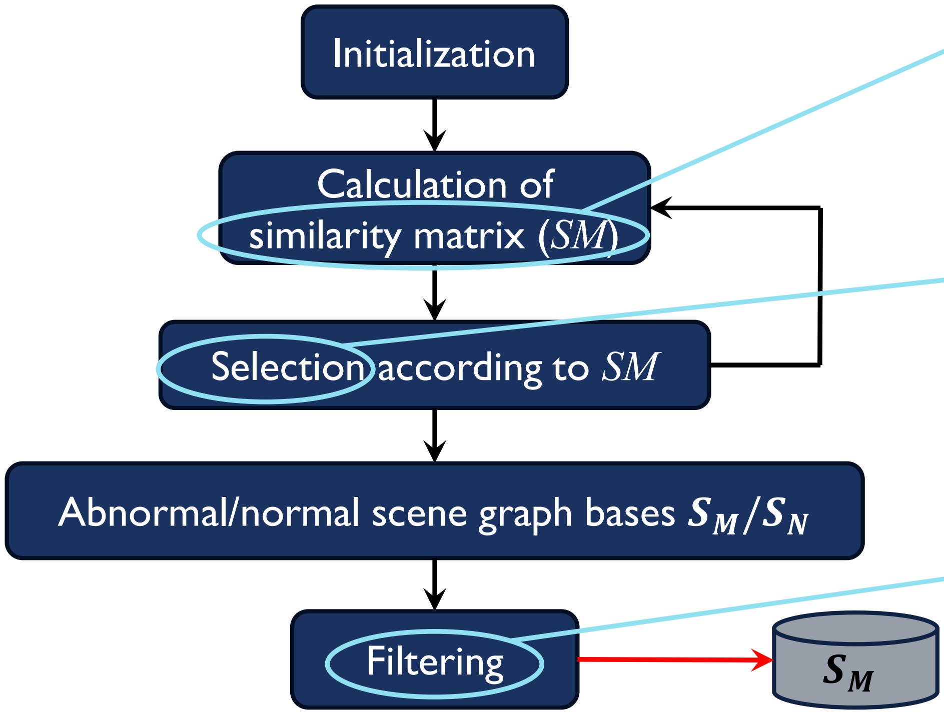
Using **graphkernel** function in Python

(2-3) Graph-based reasoning method

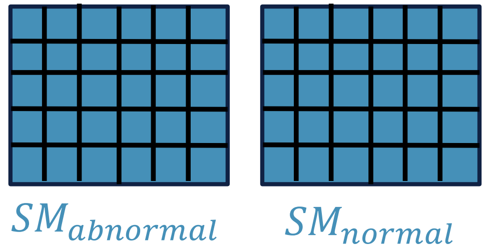


Stage1: Extraction

Objective: Extracting the abnormal scene graph base (typical abnormal samples)
Method: Clustering (DBSCAN)



SM:



Entry (i, j) : similarity value between i-th / j-th sample

Selection: for each iteration

$$S_M \leftarrow s_M^{argmax_i\, length(S_M^{neigh,i})}$$

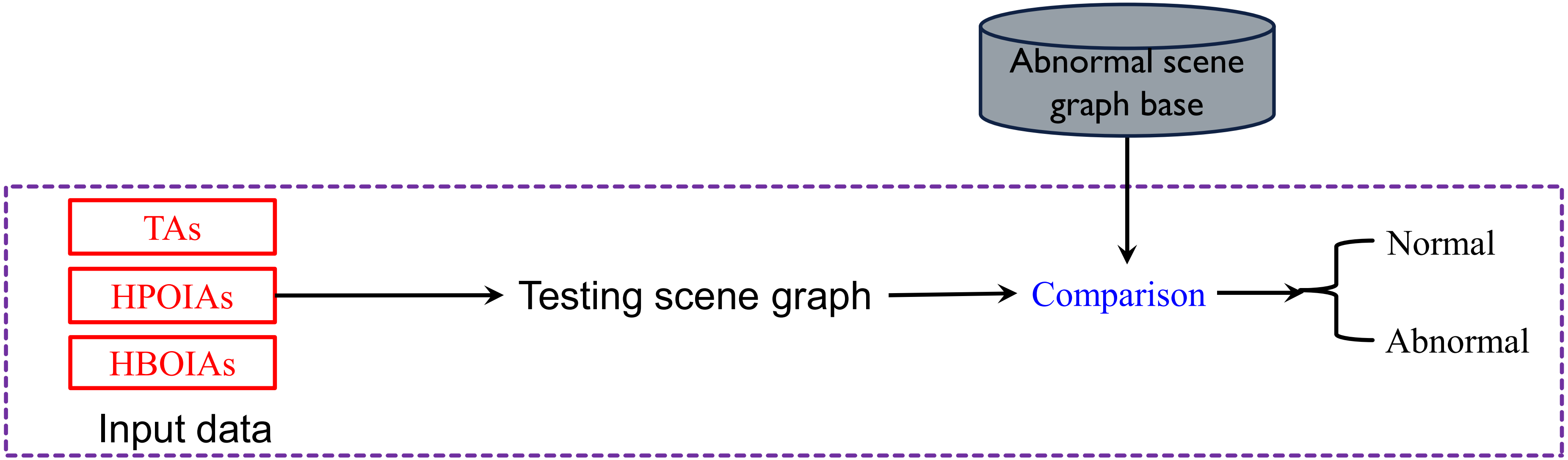
$$S_N \leftarrow s_N^{argmax_i\, length(S_N^{neigh,i})}$$

$S_M^{neigh,i}$ / $S_N^{neigh,i}$: Sample sets with high similarity to sample-i

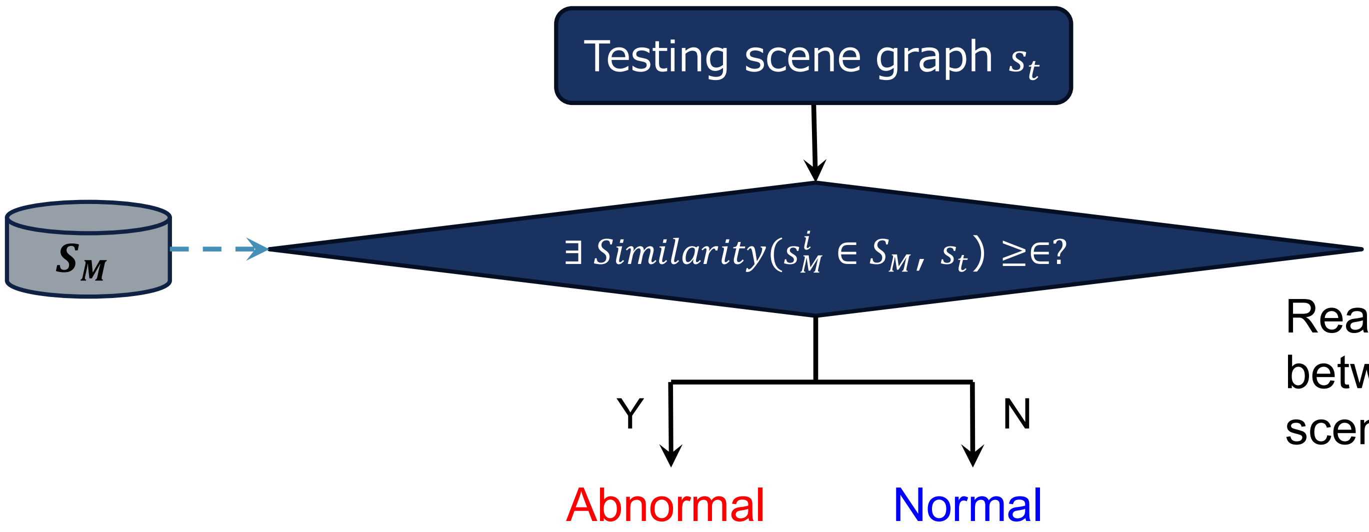
Filtering: for $s_M^i \in S_M$, if $\forall Similarity(s_M^i, s_N^j) \geq \mu$

$Delete \quad s_M^i$ **from** S_M

(2-3) Graph-based reasoning method



Stage2: Reasoning



Reasoning process is conducted between extracted $\boldsymbol{S_M}$ and testing scene graph $\boldsymbol{s_t}$

3. Experimental results

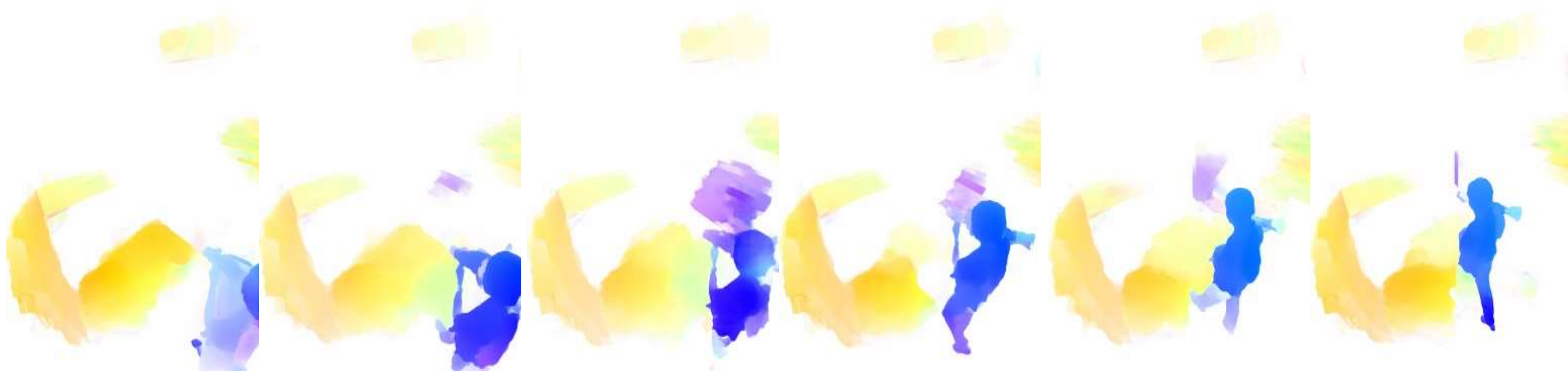
(1) Example of detection results by video-centered method

Examples of successfully detected anomalies

She holds a rocket launcher and detected as malicious.

Raw Frames

Optical Flow Frames



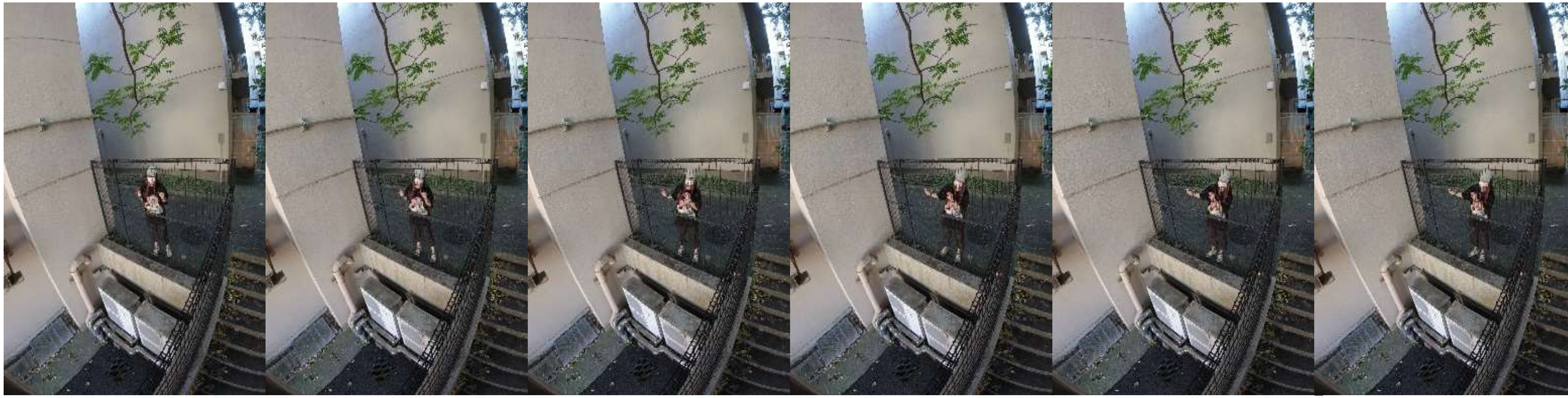
3. Experimental results

(1) Example of detection results by video-centered method

Examples of successfully detected anomalies

She climbs the wirenet !!

Raw Frames



Optical Flow Frames

3. Experimental results

(1) Example of detection results by video-centered method

Example of false detected.

She is just walking but detected as malicious.

Raw Frames

Optical Flow Frames

3. Experimental results

(1) Example of detection results by video-centered method

Example of missing of abnormality

He is cutting the wirenet !!

Raw Frames

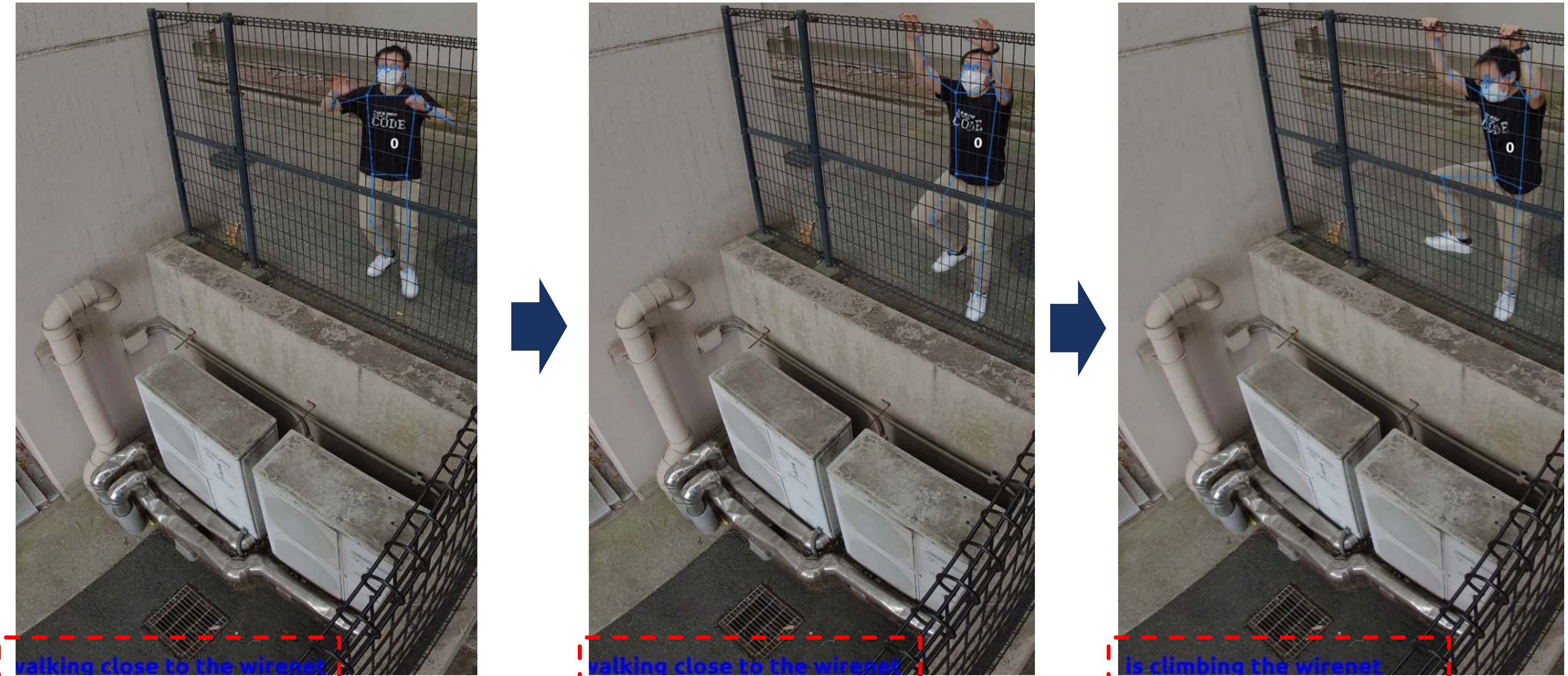
Optical Flow Frames

3. Experimental results

(2) Example of detection results by human-centered method

Examples of successfully detected anomalies

He climbs the wirenet !!



Walking close the wirenet

Walking close the wirenet

Is climbing the wirenet

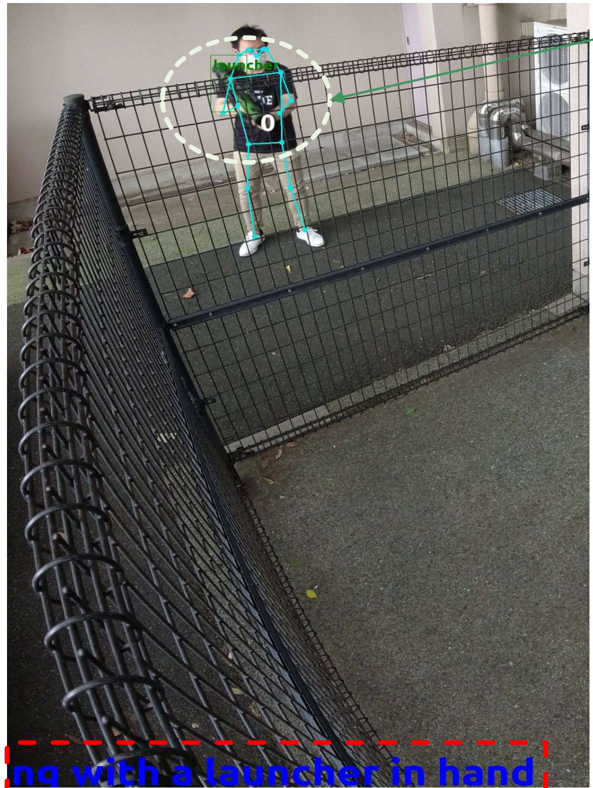
3. Experimental results

(2) Example of detection results by human-centered method

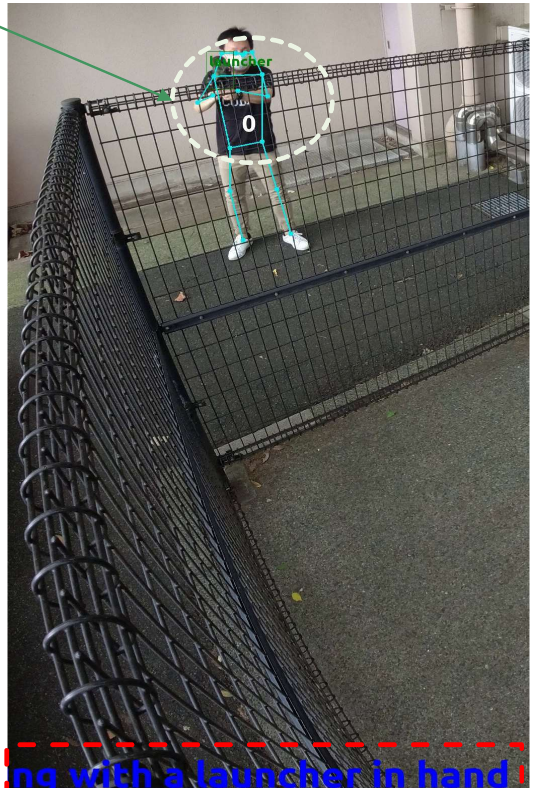
Example of missing of abnormality in a few frames

He is holding a rocket launcher

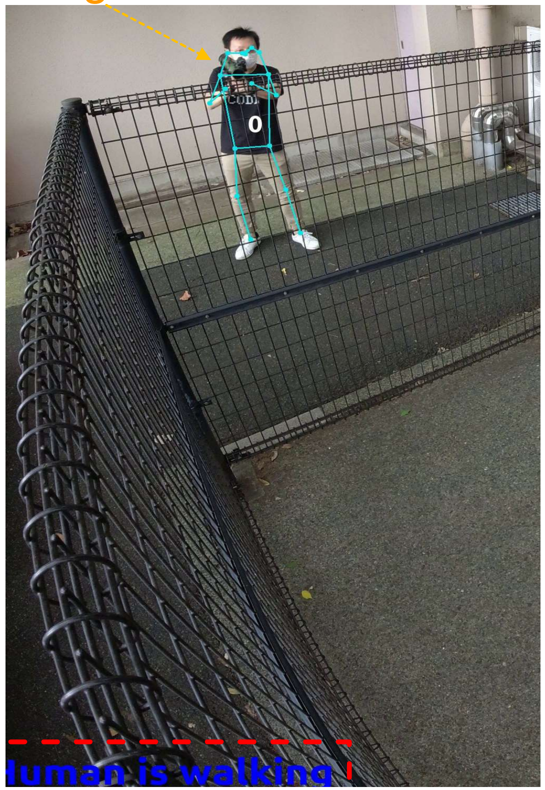
But in the next frame,
missing the rocket launcher



Holding with a launcher in hand



Holding with a launcher in hand



Human is walking

3. Experimental results of Human-centered method

(3) Comparison of video-centered and Human-centered method:

Series	Method	Precision	Recall
Video-centered	One-stream	0.7222	1.0000
	Two-stream	0.7500	0.9231
Human-centered	(Data-based)	0.7632	1.0000
	INLG	**0.7892**	**1.0000**
	JNLGI + *LMFlow*	**0.7994**	**0.9616**
	Graph-based with *JC*	0.4737	0.4737
	Graph-based with *GED*	0.5152	0.8947
	Graph-based with *GK*	0.5135	1.0000

- For more higher accuracy, plenty number of dataset for training is necessary.
- Difficult to obtain the plenty of dataset just by shooting videos.

4. GTAutoAct

It is our original framework designed to automatically generate datasets for malicious action recognition tasks.

- Rotation-orientated 3D human motion representation system.
- Coordinate transformation.
- Dynamic skeletal interpolation
- Environmental customization
- Character customization
- Map customization

5. Conclusion

- We developed **several AI model** for malicious action detection for nuclear security.
  - Video-centered method is relatively poorer than human-centered methods.
  - Language-based reasoning methods outperform others.
  - The finetune process for GPT-model is necessary.
  - Graph-based reasoning method still need to be advanced.
  - Present our task is **more detailing with high accuracy**.
- For the higher accuracy, we developed the **GTAutoAct**
  - GTAutoAct is the framework to **create** the database of video of malicious actions.
  - It has several advantages than other databases.

Thank you for your kind attention.

LI, Zhan, D2 Student

SONG, Xingyu, M2 Student

Department of Nuclear Engineering and Management,
School of Engineering, the University of Tokyo