Step 1: Check if person is wearing the required safety items

Scene Recognition Module ①

Prompt

A person wearing + item 1
⋮
A person wearing + item N

② Parse items into text prompts

Clip Text Encoder

④

Object Detection Model

③ Extract Person Image Patches

Clip Image Encoder

Pretrained VLMs

	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_1 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_M	$I_M \cdot T_1$	$I_M \cdot T_2$	$I_M \cdot T_3$	...	$I_M \cdot T_N$

⑤ **Visual-text representation**

Step 2: Check whether these items have the required specified features

Object Detection Model

⑦ **Extract Item Image Patches**

Prompt

LLM

Feature 1 + item 1
⋮
Feature N + item N

⑥ Parse items into text prompts with features

Feature-Image Matching Module

Clip Image Encoder

④

Clip Text Encoder

Pretrained VLMs

$\omega_{similarity}$

> a ?

This item has the required feature

Yes

This item doesn't have the required feature

NO

⑧ Check the similarity