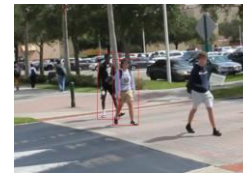
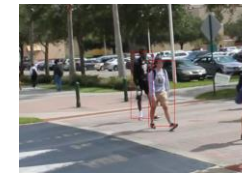
Video Frames

Object Detection and Tracking





Visual Input: Crop of two objects enclosed by bounding boxes, from two frames at time *t* and *t+30*.

Prompt: Briefly describe what the people in the enclosed regions of these images are doing. The two images were taken one second apart.

MLLM

Answer sentence r: Two people are walking side by side on the crosswalk without interacting.

Text Encoder

Embedding of r

X

Exemplar set of embeddings

Score calculation: Eq. (1) & Eq. (2)

Training mode:
If score is greater than the threshold add embedding of r to exemplar set.

Inference mode:
Use score for anomaly detection.