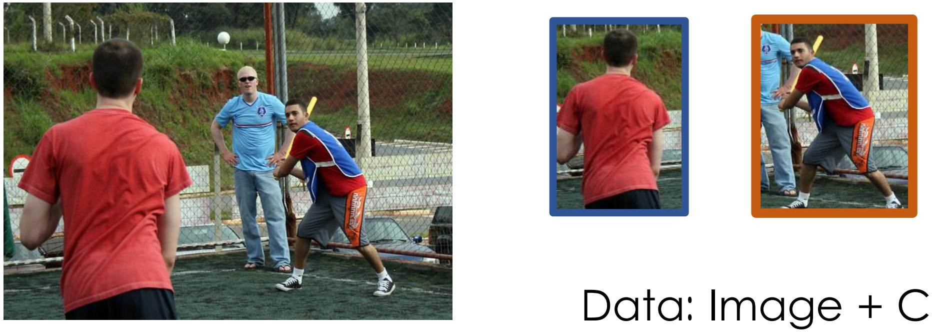
Explicit Supervision in Vision and Language Tasks

Task: Textual Grounding/Referring Expression Task:

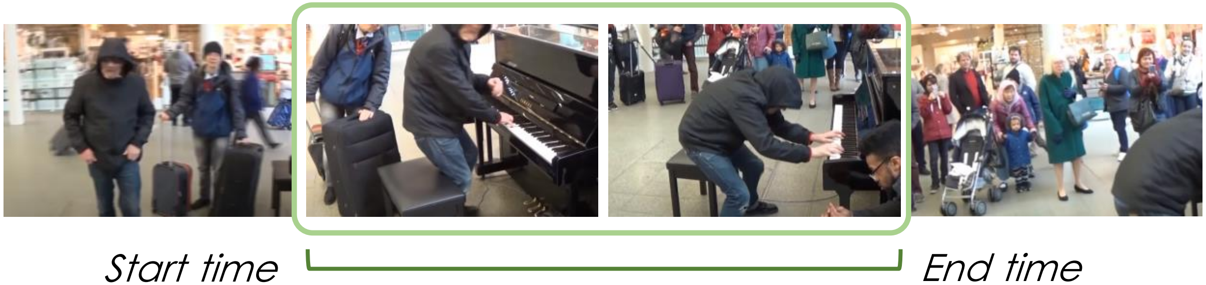


Bounding Box: (X1, Y1, X2, Y2)
"A man in red topping"

Bounding Box: (X1, Y1, X2, Y2)
"A man swing the baseball stick"

Data: Image + Caption + **Bounding Box**

Task: Video Moment Localization via Language:



Start time *End time*

"A man in black is playing the piano in the public"

Data: Video + Caption + **Temporal Boundary**