ROMA's Streaming Understanding Ability

Proactive

Event-Driven Alert

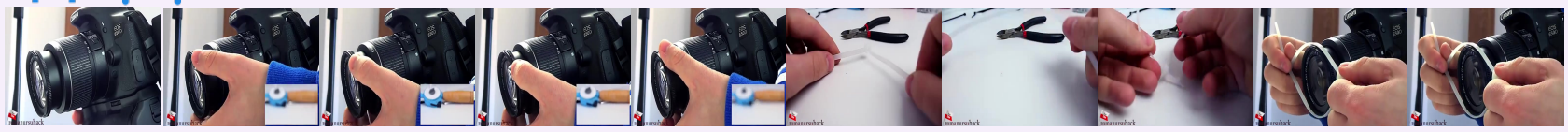
(Notify me when a bird pops out from behind a tree.)

00:00 00:01 ... 00:09 00:10 00:11 00:12 00:13 00:14 00:15 00:16

[Alert] [Alert] [Alert] [Alert]

Real-Time Narration

(What is occurring with the camera lens?)



00:00 00:01 00:02 ... 00:07 00:08 00:09 00:10 ... 00:25 00:26

One hand keeps rotating the camera lens.

Attaches the cable ties to the lens.

Reactive

Reactive QA

(What is the woman doing?) (What is the last step to complete the soup?)

00:00 00:01 ... 00:05 00:06 00:07 00:08 ... 00:18 00:19 00:20

She's cooking ... a pile of <eot>

tomatoes, then slicing them up.

Garnish the soup with chopped basil.