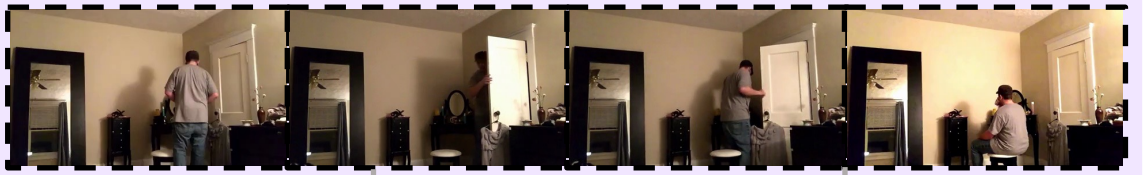
General Video Understanding

Q: Please describe this video.

A: This video depicts baseball players practicing throwing, and the presence of many teammates indicates that the sport heavily relies on teamwork.

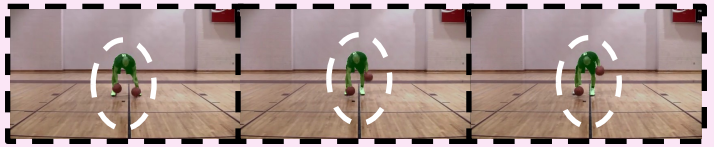
Temporal Grounding



Q: When does 'person opening a closet' happen in the video?

A: The event happen at 3.2s – 7.3s.

Video Object Referring



Q: Why might object_1<region> be focused on dribbling two basketballs simultaneously?

A: Qobject_1 may be practicing his ball-handling skills to…

Referring Video object Segmentation

Q: Please segment and track rabbit that gets food from the human.

A: The segmentation mask: [SEG].

Reasoning Video Object Segmentation.

Q: What color are the balance balls used by the athletes? Please generate a mask for the athlete on the right in each frame.

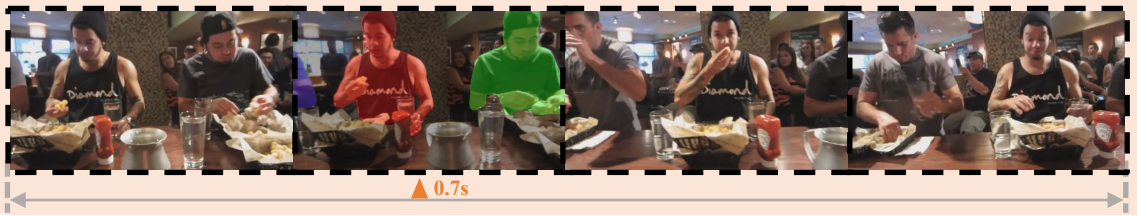
A: The balance balls used by athletes are blue. The mask: [SEG].

Pixel-level Video Referring Question Answering (PixRQA) —— *Question + Referring ⇨ Answering+ Segmentation*

Q: There are 3 objects in the video: object_1<region>, object_2<region>, object_3<region>. How does the position of object_2 change during the sequence? And please generate the mask in every frames?

A: Object_2 remains stationary beside object_1 throughout the sequence and does not move, even as object_3 kicks object_1.

Pixel-level Video Highlight Question Answering (PixHQA) —— *Question + Temporal ⇨ Answering + Segmentation*



Q: What object_1: Man wearing black beanie eating food at table, object_2: Man in gray shirt focused on eating meal, object_3: Partial view of person's shoulder and arm at table is doing in the 0.7s, and generate the mask?

A: The video features object_1, a man wearing a black beanie...Object_2, another man in a gray shirt…Object_3, a partially visible person's shoulder and arm…

Pixel-level Video Temporal Retrieval Question Answering (PixTRQA) —— *Question ⇨ Answering + Temporal + Segmentation*

Q: What object_1: Orange-handled hacksaw partially visible on right edge is doing, generate the time period and object mask?

A: The time is 3.2s – 7.3s. From left to right, there are several traditional hand...To the right of this is a small, thin-bladed hacksaw with a metallic frame and red accents. Further right, an orange-handled hacksaw (object_1) is…