We address the problem of video representation learning without human-annotated labels. While previous efforts address the problem by designing novel self-s