We invest the problem of weakly-supervised video grounding, where only video-level sentences are provided. This is a challenging task, and previous Multi-Inst