

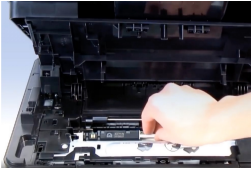







Input	VideoTaskformer (Ours)	LwDS	Correct Clip for LwDS prediction
	Take out toner cartridge	Close door of printer	
	Shake the mixture	Pour in after mixing	
	Apply detergent	Clean the floor	
	Pour noodles in water and stir	Pour cooked noodles	

Figure F1: **Step classification.** We qualitatively compare results from our method (VideoTaskformer) to the baseline LwDS on the step classification task. While the inputs are video clips, we only show a keyframe from the clip for visualization purposes. Correct predictions (VideoTaskformer) are shown in green and incorrect predictions (LwDS) are in red. We also show a frame from the clip corresponding to the incorrect prediction made by LwDS.

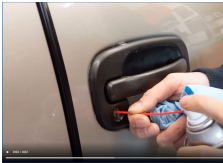
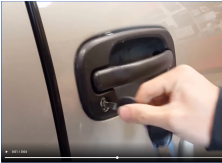




Input	Ground Truth	LwDS	VideoTaskformer (ours)
 Apply lubricant  Insert key repeatedly  Wipe off excessive lubricant Task: Lubricate A Lock	Incorrect order	Correct order	Incorrect order
 Fix the new string on the head of the guitar  Fix the new string on the lower part of the guitar  Adjust tightness of the string Task: Change Guitar Strings	Incorrect order	Correct order	Incorrect order

Figure F2: **Mistake Order Detection.** Qualitative comparison of results from VideoTaskformer to LwDS. Step and task labels shown along with the input are for visualization purpose only. Correct answers are shown in green and incorrect answers in red.