

| | | | | |
|---|--------------------------|--------------------------------|-----------------|--------------------|
| Input | | | | |
| Step labels (for visualization only) | Check type of back cover | Insert paper clip In hole | Replace battery | Install back cover |
| Step Indices | 0 | 1 | 2 | 3 |
| Ground Truth | 1 | | | |
| LwDS | 3 | | | |
| VideoTaskformer (ours) | 1 | Correct step for visualization | | |

Figure F3: **Mistake Step 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.

| Task | Input | Ground Truth | LwDS | VideoTaskformer (ours) |
|---------------------------------|---|---|---|---|
| Procedural Activity Recognition | Clean window surface Take off front of sticker Put on sticker Press sticker Tear off other side of sticker Task label: Paste car sticker | Paste car sticker | Remove scratches from windshield | Paste car sticker |
| Short-Term Step Forecasting | 1. Insert paper clip into lock 2. Twist paper clip by hand Task label: Open lock with paper clips | 3. Insert paper clip into lock | 3. Install the new doorknob | 3. Insert paper clip into lock |
| Long-Term Step Forecasting | 1. Unscrew the screws used to fix the screen Task label: Replace laptop screen | 2. Pull out screen connector, 3. Remove the screen, 4. Install new screen, 5. Reset and screw on screw | 2. Unscrew the screws, 3. Reset and screw on screw | 2. Pull out screen connector, 3. Remove the screen, 4. Install new screen, 5. Reset and screw on screw |

Figure F4: Qualitative results for **procedural activity recognition, short term step forecasting, and long term step forecasting.** 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.

Mistake Ordering Detection. Fig. F2 compares results of our method VideoTaskformer to the baseline LwDS on the mistake ordering detection task. We show two examples, “lubricate a lock” and “change guitar string”, where the steps in the input are swapped as shown by red arrows. Our method correctly detects that the input steps are in the incorrect order whereas the baseline predicts the ordering to be correct. As seen, detecting the order requires a high level understanding of the task structure, which our model learns through masking.

Mistake Step Detection. Qualitative comparison on the mistake step detection task is shown in Fig. F3. The input consists of video clip steps for the task “change battery of watch”. The second step is swapped with an incorrect step from a different task. Our method correctly identifies the index of the mistake step 1, whereas the baseline predicts 3 which is incorrect. We show the correct step for visualization purposes.

Procedural Activity Recognition. A result is shown in