Fig. F4. VideoTaskformer's representations are context-aware and can identify the right task given the sequence of clips, *"paste car sticker"*. The baseline misidentifies the task as an incorrect similar task, *"remove scratches from windshield"*.

Short-term Step Forecasting. Fig. F4 shows an input consisting of two clips corresponding to the first two steps for the task *"open lock with paper clips"*. The clips are far apart temporally, so the model needs to understand broader context of the task to predict what the next step is. Our method VideoTaskformer correctly identifies the next step as *"insert paper clip into lock"* whereas the baseline incorrectly predicts a step *"install the new doorknob"* from another task.

Long-term Step Forecasting. In Fig. F4 we compare the future steps predicted by our model and the baseline LwDS on the long-term step forecasting task. Both models only receive a single clip as input, corresponding to the first step *"unscrew the screws used to fix the screen"* of the task *"replace laptop screen"*. Our model predicts all the next 4 ground-truth steps correctly, and in the right order. The baseline on the other hand predicts steps from the same task but in the incorrect order.

All of the above qualitative results further support the effectiveness of learning step representations through masking, and show that our learned step representations are *"context-aware"* and possess *"global"* knowledge of task-structure.