is a crucial component of verifying the quality of instructional video representations. We introduce a mistake detection task and dataset for verifying if the task in a video is executed correctly—i.e. if each step is executed correctly and in the right order.

Our goal is to learn representations for the steps in the instructional video which capture semantics of the task being performed such that each step representation contains information about the surrounding context (other steps in the task). To this end, we train a model VideoTaskformer, using a masked step pre-training approach for learning step representations in instructional videos. We learn step representations jointly for a whole video, by feeding multiple steps to a transformer, and masking out a subset. The network learns to predict labels for the masked steps given just the visual representations of the remaining steps. The learned contextual representations improve performance on downstream tasks such as forecasting steps, classifying steps, and recognizing procedures.

Our approach of modeling steps further enables a new method for mistake identification. Recall, our original goal was to assist a user following an instructional video. We synthetically generate a mistakes dataset for evaluation using the step annotations in COIN [25]. We consider two mistake types: mistakes in the steps of a task, and mistakes in the ordering of the steps of a task. For the first, we randomly replace the steps in a video with steps from a similar video. For the second, we re-order the steps in a task. We show that our network is capable of detecting both mistake types and outperforms prior methods on these tasks.

Additionally, we evaluate representations learned by VideoTaskformer on three existing benchmarks: step classification, step forecasting, and procedural activity recognition on the COIN dataset. Our experiments show that learning step representation through masking pre-training objectives improves the performance on the downstream tasks. We will release code, models, and the mistake detection dataset and benchmark to the community.

2. Related Works

Instructional Video Datasets and Tasks. Large-scale narrated instructional video datasets [6, 17, 25, 30, 31] have paved the way for learning joint video-language representations and task structure from videos. More recently, datasets such as Assembly-101 dataset [21] and Ikea ASM [3] provide videos of people assembling and disassembling toys and furniture. Assembly-101 also contains annotations for detecting mistakes in the video. Some existing benchmarks for evaluating representations learned on instructional video datasets include step localization in videos [6, 25], step classification [6, 25, 31], procedural activity recognition [25], and step forecasting [13]. In our work, we focus on a broad range of instructional videos found in HowTo100M [17] and evaluate the learned representations on the downstream tasks in COIN [25] dataset. We additionally introduce 3 new benchmarks for detecting mistakes in instructional videos and forecasting long-term activities.

Procedure Learning from Instructional Videos. Recent works have attempted to learn procedures from instructional videos [2, 5, 13, 19, 27]. Most notably, [5] generates a sequence of actions given a start and a goal image. [2] finds temporal correspondences between key steps across multiple videos while [19] distinguishes pairs of videos performing the same sequence of actions from negative ones. [13] uses distant supervision from WikiHow to localize steps in instructional videos. Contrary to prior works, our step representations are aware of the task structure as we learn representations globally for all steps in a video jointly, as opposed to locally, as done in past works.

Video Representation Learning. There has been significant improvement in video action recognition models over the last few years [1, 9, 10, 14]. All of the above methods look at trimmed videos and focus on learning short-range atomic actions. In this work, we build a model that can learn longer and more complex actions, or steps, composed of multiple short-range actions. For example, the first step in Fig. 1, *"Make batter"*, is composed of several atomic actions such as *"pour flour"* and *"whisk"*. There have also been works [13, 16, 20, 23, 29] which learn representations for longer video clips containing semantically more complex actions. Our work falls into this line of work.

3. Learning Task Structure through Masked Modeling of Steps

Our goal is to learn task-aware step representations from a large corpus of instructional videos. To this end, we develop VideoTaskformer, a video model pre-trained using a BERT [7] style masked modeling loss. In contrast to BERT and VideoBERT [23], we perform masking at the step level, which encourages the network to learn step embeddings that encapsulate the semantics and temporal ordering of steps within the task.

Our framework consists of two steps: pre-training and fine-tuning. During pre-training, VideoTaskformer is trained on weakly labeled data on the pre-training task. For fine-tuning, VideoTaskformer is first initialized with the pre-trained parameters, and a subset of the parameters is fine-tuned using labeled data from the downstream tasks. Each downstream task yields a separate fine-tuned model.

We first provide an overview of the pre-training approach before delving into details of the individual components.

Overview. Our approach for pre-training VideoTaskformer is outlined in Fig. 2. Consider an instructional video V consisting of K video clips $v_i, i \in [1, \ldots, K]$ corresponding to K steps in the video. A step $v_i \in \mathbb{R}^{L \times H \times W \times 3}$ is a sequence of L consecutive frames depicting a step, or semantic com-