GAIN: ON THE GENERALIZATION OF INSTRUCTIONAL ACTION UNDERSTANDING

Junlong Li[1], **Guangyi Chen**[2,3], **Yansong Tang**[1], **Jinan Bao**[4],
Kun Zhang[2,3], **Jie Zhou**[1], **Jiwen Lu**[1,*]
[1]Tsinghua University, [2]MBZUAI, [3]Carnegie Mellon University, [4] University of Alberta

ABSTRACT

Despite the great success achieved in instructional action understanding by deep learning and mountainous data, deploying trained models to the unseen environment still remains a great challenge, since it requires strong generalizability of models from in-distribution training data to out-of-distribution (OOD) data. In this paper, we introduce a benchmark, named **GAIN**, to analyze the Generaliz-Ability of INstructional action understanding models. In GAIN, we reassemble steps of existing instructional video training datasets to construct the OOD tasks and then collect the corresponding videos. We evaluate the generalizability of models trained on in-distribution datasets with the performance on OOD videos and observe a significant performance drop. We further propose a simple yet effective approach, which cuts off the excessive contextual dependency of action steps by performing causal inference, to provide a potential direction for enhancing the OOD generalizability. In the experiments, we show that this simple approach can improve several baselines on both instructional action segmentation and detection tasks. We expect the introduction of the GAIN dataset will promote future in-depth research on the generalization of instructional video understanding. The project page is https://jun-long-li.github.io/GAIN.

1 INTRODUCTION

Instructional videos play an essential role for learners to acquire different tasks. The explosion of instructional video data on the Internet paves the way for learners to acquire knowledge and for computer vision community training models, for example, human can train an action segmentation model to understand the video by the dense step prediction of each frame, or an action detection model to localize each step. While a number of datasets for instructional action understanding (IAU) have been proposed over the past years(Alayrac et al., 2016; Das et al., 2013b; Malmaud et al., 2015; Sener et al., 2015) and growing efforts have been devoted to learning IAU models(Zhukov et al., 2019; Huang et al., 2017), the limited generalizability of models remains to be a major obstacle to the deployment in real-world environments. One may ask a question "*Suppose the model has learned how to inflate bicycle tires, does it know how to inflate car tires?*" In fact, due to potential environmental bias between the training dataset and application scenes, the well-trained model might not be well deployed in an OOD environment (Ren et al., 2019), especially when instructional videos of interest to users are not involved in the finite training dataset.

To encourage models to learn transferable knowledge, it is desirable to benchmark their generalizability. Though this OOD generalization problem (Barbu et al., 2019; Hendrycks et al., 2021; Hendrycks & Dietterich, 2019) attracts much attention in the field of image recognition, such as ObjectNet (Barbu et al., 2019) and ImageNet-R (Hendrycks et al., 2020), it has barely been explored for the IAU task. A related problem is video domain generalization (Yao et al., 2021) (VDG) for conventional action recognition which focuses on domain generalization when changing the scene or background of the action. However, different from conventional action, the key obstacle to the generalization of instructional action is the distribution shift of action steps under different task categories, which is caused by the collection bias of the datasets. In Fig. 3, we show that the steps under different task categories have different distributions.

*Corresponding author