Exploring Ordinal Bias in Action Recognition for Instructional Videos

Anonymous authors
Paper under double-blind review

Abstract

Action recognition models have shown promising results in understanding consecutive human actions in instructional videos. However, they often rely on dominant action patterns in datasets rather than achieving true video comprehension. We define this as ordinal bias, a systematic reliance on dataset-specific action sequences. To mitigate this, we introduce two simple yet effective video manipulation techniques: action masking and sequence shuffling, where the latter action in dominant pairs is masked, or the sequence is randomized. Our findings reveal that existing models still tend to rely on dominant action pairs and struggle to adapt, highlighting their overestimated performance and lack of robustness.

1 Introduction

Action recognition in instructional videos has witnessed remarkable progress, primarily driven by models that excel in curated benchmark datasets (Farha & Gall, 2019; Ishikawa et al., 2021; Li et al., 2020; Yi et al., 2021). However, these datasets often present a limited view of real-world variability by favoring specific, repeated action sequences. In realistic settings, such as home surveillance, autonomous driving, or user-generated content on social media, actions occur in an unpredictable and nonstandard order. As a result, models trained on conventional benchmarks tend to exploit these spurious ordinal patterns, a phenomenon that we refer to as ***ordinal bias***.

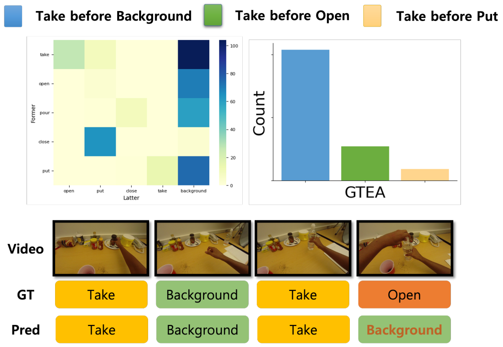

Figure 1: **Illustration of the ordinal bias.** The model incorrectly predicts 'Background' instead of 'Open', likely because the 'Take-Background' pair is dominant.

We observe that existing datasets (Fathi et al., 2011; Stein & McKenna, 2013; Kuehne et al., 2014) demonstrate biased action sequences, which leads the model to suffer from spurious correlations. As shown in Figure 1, the dataset exhibits a dominant occurrence of the action 'Take' followed by 'Background.' This biases the model toward learning spurious correlations, causing it to predict 'Background' as the next action rather than relying on visual inputs to correctly predict 'Open.' This raises concerns about the reliability of evaluations and the risk of overestimating the performance of the model.

To address this, we propose two video manipulation methods: ***Action Masking*** and ***Sequence Shuffling*** for a reliable evaluation. In the action masking method, we selectively mask or replace the video frames corresponding to a specific action unit with a 'no action' label, compelling the model to depend on alternative contextual visual cues rather than on learned ordinal patterns. In contrast, the sequence shuffling method randomly rearranges the order of the action labels while keeping the frame order within each action unit intact.

With our methods, our experiments reveal that state-of-the-art action recognition models struggle to generalize manipulated videos, demonstrating their lack of robustness. Furthermore, even when models are trained on videos with mitigated action distributions through our manipulation techniques, they still tend to capture dominant action pairs in datasets. These findings highlight the pressing need to rethink an evaluation framework, a training strategy, and advanced modeling so that models can adapt to real-world scenarios.

2 Related Work

2.1 Recognition of Actions in Instructional Videos

Instructional video analysis has emerged as a prominent area of research in the field of video comprehension. In particular, multiple datasets of instructional videos (Fathi et al., 2011; Stein & McKenna, 2013; Kuehne et al., 2014) have been introduced, offering extensive contextual information on human activities. Despite several proposed techniques (Farha & Gall, 2019; Li et al., 2020; Ishikawa et al., 2021; Yi et al., 2021; Li et al., 2022; Liu et al., 2023) for these datasets, the ordinal bias issue can overestimate their performance. This comes from their exploitation of the action sequence observed during training. Recent work such as Liu et al. (2022) has started to address ordinal action understanding by explicitly modeling inter-action context in instructional videos.

2.2 Bias in Action Recognition

In recent years, several studies (Li et al., 2018; Nam et al., 2020; Hara et al., 2021; Duan et al., 2023) have investigated the issue of bias in action recognition. A line of works (Duan et al., 2022; Zhai et al., 2023; Li et al., 2024) have explored the dual challenges of background and foreground biases, demonstrating that action recognition models can be inadvertently biased by static and dynamic cues. Duan et al. (2023) showed that the incorporation of adversarial losses can help reduce bias in action representations. These advances mark significant progress in the development of fair and robust action recognition systems. Although previous work such as Nam et al. (2020) addresses bias by retraining classifiers on the errors of biased models, a process that can be computationally expensive and still overlooks the underlying imbalance of the data set, our approach directly addresses the issue by manipulating the video data itself. Furthermore, we provide insights into how current models perform if they are trained with these video variants.

3 Experimental Setup

In this section, we first introduce three video datasets and five action recognition models for our experiment. Then we provide evaluation metrics with details.

3.1 Dataset

We utilize three action recognition datasets: Georgia Tech Egocentric Activities (GTEA) (Fathi et al., 2011), 50Salads (Stein & McKenna, 2013), and Breakfast (Kuehne et al., 2014). GTEA includes 28 videos that depict daily kitchen activities, featuring 11 action categories. Each video has an average of 20 action units and a duration of approximately 30 seconds. 50Salads contains 50 videos of actors preparing salads in various kitchen environments, with more than 20 actors participating. The videos in 50Salads are more than six minutes long and cover 17 action categories. Breakfast consists of over 1700 videos that contain breakfast preparation scenes and has 48 action categories. This dataset has the most complex labeling scheme among the three datasets.

3.2 Action Recognition Models

We consider five distinct models: MS-TCN (Farha & Gall, 2019), MS-TCN++ (Li et al., 2020), ASRF (Ishikawa et al., 2021), and DiffAct (Liu et al., 2023). MS-TCN employs a multi-stage architecture with dilated temporal convolutions and a smoothing loss to iteratively refine frame-level predictions. MS-TCN++ extends this approach by integrating dual dilated layers that capture both local and global contexts while decoupling prediction generation from refinement. ASRF improves segmentation quality by adding an auxiliary branch that explicitly regresses action boundaries to mitigate over-segmentation errors. ASFormer leverages a Transformer-based framework augmented with temporal convolutions and a hierarchical representation pattern for iterative prediction refinement. Lastly, DiffAct formulates action segmentation as a conditional sequence generation task that

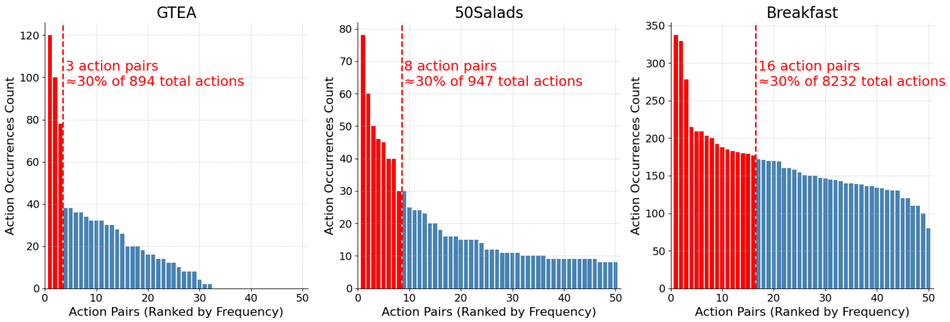

Figure 2: **Long-tailed distributions of action pairs in datasets.** Each dataset's histogram represents the frequency of action pairs, ranked by their occurrence count. The red-highlighted action pairs contribute to 30% of all actions in the dataset, despite being a small fraction of the total pairs.

iteratively denoises a noisy action sequence by leveraging priors such as positional, boundary, and relational cues. In all our experiments, we utilized I3D (Carreira & Zisserman, 2017) video features, which were pre-trained using the Kinetics dataset (Kay et al., 2017), and a single NVIDIA RTX 3090.

3.3 Evaluation Metric

To evaluate the performance of the model, we utilize frame-wise accuracy, a primary metric that gauges the percentage of accurately classified actions within a unit frame of a test dataset. To produce a result, we use 5-fold cross-validation to evaluate the proposed approach's performance on the 50Salads dataset. For the remaining datasets, 4-fold cross-validation is performed to estimate the average performance measure.

4 Ordinal Bias Problem

4.1 Long tail distribution of action pairs

We begin by analyzing the distribution of action pairs across the datasets. As shown in Figure 2, each dataset exhibits a pronounced long-tailed pattern. In detail, in the Breakfast dataset, 16 out of 228 action pairs represent 30% of all action occurrences. Similarly, in the 50Salads dataset, 8 out of 120 action pairs contribute to 30% of the total action occurrences, and in the GTEA dataset, only 3 out of 32 action pairs comprise 30% of the overall instances. This skewed distribution may lead to biased predictions, as models can become overly influenced by the few frequently occurring action pairs, potentially misrepresenting the diversity of real-world actions. To address this issue and enable more reliable evaluations, we introduce video manipulation methods designed to counteract the effects of this long-tailed distribution.

4.2 Video Manipulation Methodologies

We propose two video manipulation techniques, Action Masking and Sequence Shuffling, as shown in Figure 3, For the action masking, we mask the video frames of a specific action unit, and the corresponding action label is replaced with ‘no action.’ By doing so, we verify whether the model predicts ‘no action’ accounting for visual variants or if it makes biased predictions. The sequence shuffling randomly rearranges the order of action without changing the order of the frames within each action unit. This technique allows us to construct a dataset with a distinctive label distribution from the original, thereby mitigating the presence of skewed distribution and ensuring reliable evaluations.

4.3 Evaluation with Proposed Manipulation Methods

Figure 4 shows result of our methods, demonstrating that our manipulation methods successfully change the distributions of action pairs. Specifically, for action masking, all the subsequent labels of

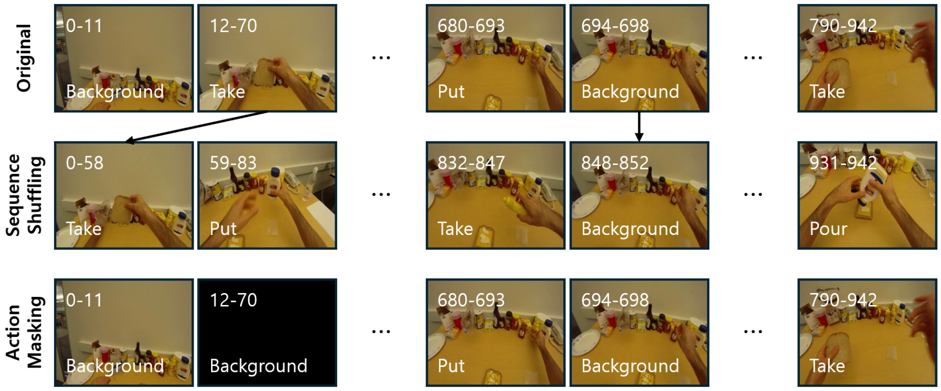

Figure 3: **Manipulation techniques**. Each video contains 943 frames. A single image represents consecutive frames, displayed in the top-left corner, while the action label is in the bottom-left corner. In the sequence shuffling, frames are shuffled in sequences, e.g., moving frames 12 to 70 to positions 0 to 58, and frames 695 to 698 to positions 848 to 852. In action masking, frames 12 to 70 are masked and labeled as background.

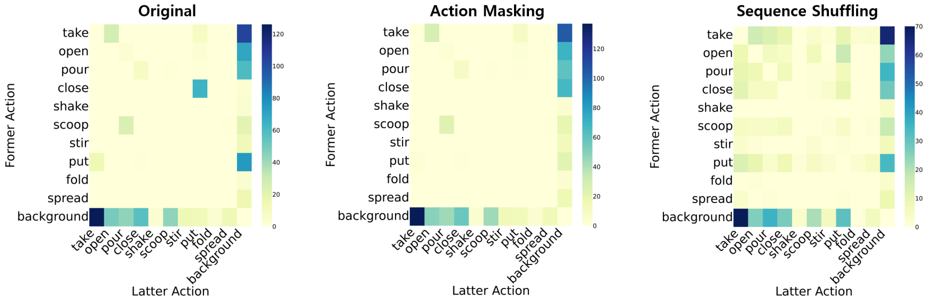

Figure 4: **Heatmap of the frequency of action pairs with GTEA dataset**. We use the initial action ‘close’ for action masking.

‘close’ have been switched to ‘background’, whereas the sequence shuffling reduced the maximum value of high occurrences and introduced a new pair of actions. Also, in the sequence shuffling, the number of existing biased pairs is decreased while previously absent action pairs are created. More visual examples can be found in Appendix D.1

Now, we apply action masking technique to the original dataset and conduct an experiment to see how the model trained with the original dataset behaves when it encounters a masked section. We first select the action pair that is frequently seen in the original dataset. We selected action pairs according to these criteria: 1) Considering the initial action, we observe the frequency of subsequent actions to determine if a particular combination is significantly more common compared to the others. 2) Subsequent actions should not equate to ‘no action’. As a result, in the GTEA dataset, ‘close’ is used as a prior action, comprising 7.5% of the entire dataset, with the ‘put’ combination comprising 95.5% of these actions. Similarly, in the Breakfast dataset, ‘pour_dough2pan’ serves as the initial action, accounting for 1.8% of the total dataset, while ‘fry pancake’ constitutes 91.1% of these activities. Then, we mask frames that correspond to the latter action unit and replace its action label with ‘no action.’ Lastly, we make the model predict the masked parts and inspect the accuracy.

Figure 5 shows the results of our experiment, demonstrating that the model finds it difficult to accurately predict the manipulated test videos. This result indicates that the model misclassifies masked regions as having an action label from the original dataset instead of identifying them as ‘no action,’ suffering from the ordinal bias problem. This also indicates that the model does not utilize visual information, but exploits spurious correlation for prediction. We will discuss the ordinal bias problem in detail in the next section.

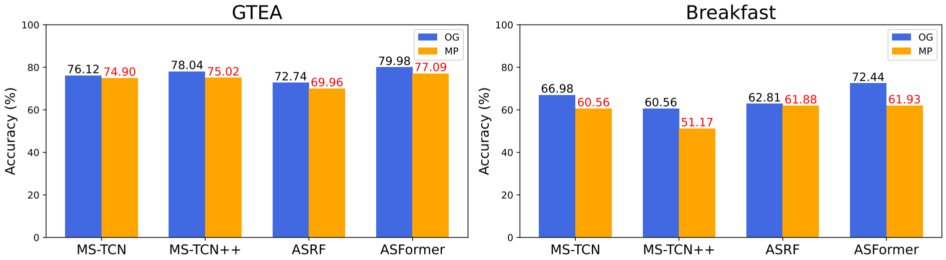

Figure 5: **The results on the original (OG) and manipulated (MP) test set**. Consistent performance drops across different datasets and models suggest that current models rely on the ordinal bias in the datasets.

5 Analysis of Ordinal Bias Problem

Within this section, we explore how models contribute to the issue of ordinal bias.

5.1 Evaluation of Model Generalization

We investigate the extent to which a model is responsible for the problem by training it on a manipulated dataset using the sequence shuffling method described in Section 4.2. Here, we skip the action masking as it only introduces no-action labels, therefore specific actions are only followed by 'no-action.' In contrast, sequence shuffling provides more diverse action pairs, allowing us to assess how models handle varied action patterns. We then evaluate the performance of the model on the original dataset and sequence shuffling. A model with satisfactory generalization should exhibit good performance in the original dataset despite being trained on a manipulated one. Table 1 shows a significant discrepancy in performance between the model trained on the modified dataset (OG) and the superior performance of the model trained on the unchanged dataset (OG*).

Furthermore, we have investigated whether these results come from ordinal bias by comparing the label distribution among the original dataset, the manipulated dataset, and the model's predictions. If the model is robust, its prediction distribution (green) should resemble the manipulated dataset distribution (blue), rather than the original dataset distribution (red). Figure 6 shows results, which implies that the model tends to make prediction by following the trend of the training data set, not by given visual information. This outcome implies that the model exploits spurious correlations during inference to achieve higher scores, resulting in an overestimation. Therefore, a model must have an improved generalization capability to reduce the ordinal bias.

Dataset	MS-TCN			MS-TCN++			ASRF			ASFormer			DiffAct		
	OG*	OG	SH	OG*	OG	SH	OG*	OG	SH	OG*	OG	SH	OG*	OG	SH
GTEA	76.12	64.50	69.63	78.04	65.08	67.52	72.74	58.08	69.46	79.98	71.61	76.80	80.30	72.68	76.89
Breakfast	66.98	50.08	56.13	60.56	47.96	54.77	62.81	54.27	57.30	72.44	–	–	76.59	–	–
50Salads	79.33	69.85	71.17	74.89	72.54	73.43	82.14	65.13	70.69	85.62	65.41	66.93	88.43	76.20	78.39

Table 1: **Accuracy of the models trained on the sequence shuffling dataset. OG***: model trained and tested on the original dataset. **OG**: model trained on the sequence shuffling dataset and tested on the original test set. **SH**: model trained on the sequence shuffling dataset and tested on its test set. Due to memory constraints, we were not able to test DiffAct and ASFormer on Breakfast.

5.2 Impact of Additional Training

In many bias-related problems, the incorporation of additional data helps alleviate bias. Therefore, we investigate whether training models with an additional augmented dataset can mitigate ordinal bias in action recognition. To this end, we designed a curriculum learning-like strategy by sequentially training the model on three variants of the dataset: the original, the masked, and the shuffled

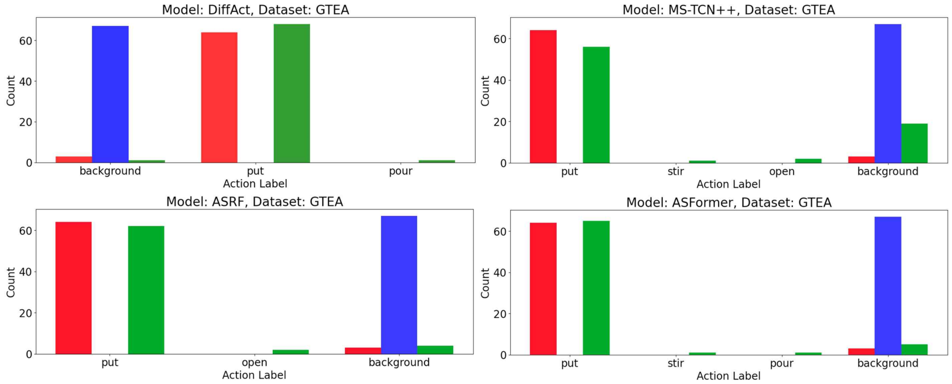

Figure 6: **Distribution of the action labels on GTEA dataset**. Red: distribution of original dataset label. Blue: distribution of action masking dataset label. Green: distribution of model predicted label. For more visualization, please refer to the Appendix D.2.

versions. This progression is intended to gradually expose the model to increasing difficulty levels and reduce its reliance on spurious correlations. However, as highlighted in Section 5.1, the use of action masking is not suitable for model training. To overcome this, we used a strategy for action masking inspired by Masked Language Modeling (Devlin, 2018), where actions were randomly hidden with a likelihood 15%, instead of our original approach. In contrast, for sequence shuffling, we utilized the method we had proposed. Consequently, every augmented dataset retains its original size, allowing the model to be trained on a dataset that is triple the size of the initial one.

However, as shown in Table 2, the model trained with an additional dataset did not exceed the performance of the model trained solely on the original dataset. These results indicate that simply augmenting the training data, even through a curriculum-learning-like approach, does not effectively mitigate ordinal bias. This suggests that the bias is deeply ingrained in the training dynamics, and additional intervention, such as architectural modifications or specialized loss functions, may be required to address the issue. The ablation study can be found in the Appendix B.

Dataset	MS-TCN++				ASFormer			
	OG	MP	S	M	OG	MP	S	M
GTEA	78.04	70.28	69.20	75.77	79.98	76.91	72.80	77.91
Breakfast	60.56	50.75	49.41	46.42	72.44	-	-	-

Table 2: **Accuracy of the models trained with additional datasets. OG**: Trained and tested on the original dataset. **MP**: Tested on the original dataset. **S**: Tested on the sequence shuffling dataset. **M**: Tested on the dataset using action masking. Except for OG, all models are trained on the original dataset supplemented with two additional datasets.

6 Conclusion

Our investigation of ordinal bias reveals a critical oversight in current action recognition research: the overreliance on fixed, dataset-specific action sequences. Although high accuracies are reported on popular benchmarks, such performance does not necessarily translate into reliable predictions in real-world settings, where the sequence of actions is highly variable and unpredictable. By applying our proposed video manipulation techniques, we demonstrate that models vulnerable to ordinal bias exhibit significant drops in performance when faced with non-standard action orders.

The practical implications of this work are substantial. Ensuring robust action recognition under diverse conditions is essential for applications ranging from automated surveillance to assistive robotics. Therefore, we advocate for future research to incorporate evaluation protocols that account for ordinal variations and to develop training methodologies that reduce the reliance on spurious correlations. Addressing the ordinal bias problem is vital for advancing academic research and bridging the gap between controlled experimental setups and the complexities of real-world environments.

REFERENCES

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Haodong Duan, Yue Zhao, Kai Chen, et al. Omnidebias: Leveraging web data for joint training to neutralize intrinsic dataset bias in action recognition. In *Proceedings of CVPR*, 2022.

Haodong Duan, Yue Zhao, Kai Chen, Yuanjun Xiong, and Dahua Lin. Mitigating representation bias in action recognition: Algorithms and benchmarks. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pp. 557–575. Springer, 2023.

Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3575–3584, 2019.

Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pp. 3281–3288. IEEE, 2011.

Kensho Hara, Yuchi Ishikawa, and Hirokatsu Kataoka. Rethinking training data for mitigating representation biases in action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3349–3353, 2021.

Yuchi Ishikawa, Seito Kasai, Yoshimitsu Aoki, and Hirokatsu Kataoka. Alleviating over-segmentation errors by detecting action boundaries. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2322–2331, 2021.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 780–787, 2014.

Jian Li et al. Fair action recognition: Balancing accuracy and equity in spatiotemporal models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. In press.

Muheng Li, Lei Chen, Yueqi Duan, Zhilan Hu, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Bridge-prompt: Towards ordinal action understanding in instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19880–19889, 2022.

Shi-Jie Li, Yazan AbuFarha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020. doi: 10.1109/TPAMI.2020.3021756.

Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 513–528, 2018.

Daochang Liu, Qiyue Li, Anh-Dung Dinh, Tingting Jiang, Mubarak Shah, and Chang Xu. Diffusion action segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10139–10149, 2023.

Muheng Liu, Lei Chen, Yueqi Duan, et al. Bridge-prompt: Towards ordinal action understanding in instructional videos. *arXiv preprint arXiv:2203.14104*, 2022.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020.

Sebastian Stein and Stephen J McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pp. 729–738, 2013.

Fangqiu Yi, Hongyu Wen, and Tingting Jiang. Asformer: Transformer for action segmentation. *arXiv preprint arXiv:2110.08568*, 2021.

Yuanhao Zhai, Ziyi Liu, Zhenyu Wu, et al. Soar: Scene-debiasing open-set action recognition. In *arXiv preprint arXiv:2309.01265*, 2023.

A Evaluation of Action Segmentation Performance

This section evaluates the capacity of the model to carry out the task of action segmentation. Our model demonstrates proficiency in both action segmentation and recognition tasks. We conduct an experiment on how altering instructional videos affects the model's performance in the action segmentation task by using the Shuffle method. We begin by hypothesizing that the model's effectiveness on altered videos will be comparable to or exceed that on the original videos, due to the distinctly unnatural nature of action transitions.

However, the results presented in Table A1 indicate that the performance of the ASFormer model on the manipulated video is inferior to that of the original video. These outcomes suggest that the action segmentation task may be influenced by ordinal bias, a matter that we will explore further in our research.

Metric	GTEA		Breakfast		50Salads	
	OG	MP	OG	MP	OG	MP
$Edit$	84.04	50.71	54.01	39.09	73.50	37.17
$F1@10$	88.69	59.52	61.09	48.65	74.03	40.49
$F1@25$	87.76	57.51	58.15	46.09	68.69	36.63
$F1@50$	79.02	46.27	51.33	40.05	55.02	27.31

Table A1: **Performance of ASFormer model on the action segmentation task for various datasets. OG**: performance on original datasets. **MP**: performance on manipulated dataset. Metrics include segmental edit score and segmental overlap $F1$ score at a threshold of $k/100$ where k equals the percentage of overlap, denoted as $F@k$.

B Ablation Experiment of Additional Training

This section reports on the ablation study in an additional dataset. As exhibited in Table A2, each outcome falls short of the initial performance, indicating that further training might not address the ordinal bias issue. Note that, as 50Salads does not have a label that refers to 'no-action', we omit the results that use the action masking method. Also, we have not presented the results for the Breakfast dataset when using ASFormer and DiffAct due to an inability to replicate these results.

C Revisiting Video Manipulation Method

Proposed manipulation technique is effective in judging whether the model utilizes visual cues well or not. However, this methodology could lead to the following problems. For the action masking method, masked part of the frame may represent inaccurate inferences because they may be parts of the frame that the model did not encounter during training. Also, the sequence shuffling produces quite unnatural video context, as we randomly shuffle sequence of actions.

To complement this issue, we use a the sequence shuffling technique, but rather than shuffling randomly, we replace the latter action in frequently occurring action pairs with a random action and

Data	Method	GTEA		Breakfast		50Salads	
		OG	MP	OG	MP	OG	MP
M+S	DiffAct	80.30	78.86	-	-	-	-
	MS-TCN++	78.04	72.79	60.56	54.69	-	-
	ASFormer	79.98	73.64	-	-	-	-
O+M	DiffAct	80.30	78.00	-	-	-	-
	MS-TCN++	78.04	75.39	60.56	59.73	-	-
	ASFormer	79.98	79.39	-	-	-	-
O+S	DiffAct	80.30	74.80	-	-	88.43	65.43
	MS-TCN++	78.04	69.78	60.56	54.85	74.89	72.79
	ASFormer	79.98	76.61	-	-	85.61	77.43

Table A2: **Result of ablation study on additional dataset O**: Original. **S**: the sequence shuffling. **M**: Action Masking. **OG**: Model trained on original dataset. **MP**: Model trained on additional dataset.

location within the video. This technique will henceforth be referred to as 'Limited Shuffling.' Table A3 shows experiment results, still revealing that the model suffers from an ordinal bias problem. For qualitative results, refer to Section D.3.

	DiffAct		MS-TCN++		ASRF		ASFormer	
	OG	MP	OG	MP	OG	MP	OG	MP
GTEA	80.30	71.94	78.04	70.97	72.74	67.20	79.98	71.90
Breakfast	76.59	74.32	60.56	58.98	62.81	60.66	72.44	70.44
50Salads	88.43	82.56	74.89	69.37	82.14	74.16	85.61	70.07

Table A3: **Accuracy of model evaluated on Limited Shuffling method. OG**: performance on the original dataset **MP**: performance on limited shuffling method.

D Detailed Qualitative Results

This section displays the visualization results mentioned in the main paper.

D.1 Visualization of Action Pair Distribution

Figure A1 illustrates the visualization of the frequency of the pair of action labels of 2 grams within the Breakfast dataset using the action Masking method. Furthermore, Figure A2 presents results from the 50Salads dataset with the Shuffle Dataset approach, while Figure A3 shows results from the Breakfast dataset also employing the Shuffle Dataset technique.

D.2 Qualitative results of Model Prediction on Action Masking

Figure A4 shows the distribution of predicted and ground truth labels in breakfast with the applied action masking technique of the data set. '(pour_dough2pan, fry_pancake)' pair was used for the result.

D.3 Qualitative results of Model Prediction on Limited Shuffling

Figure A5, Figure A6, Figure A7, and Figure A8 show the distribution of the original label, limited shuffling label, and the prediction of the model in the limited shuffling dataset. The '(close, put)' action pair is selected for GTEA, '(cut_tomato, place_tomato_into_bowl)' for 50Salads, and '(pour_dough2pan, fry_pancake)' for Breakfast, respectively.

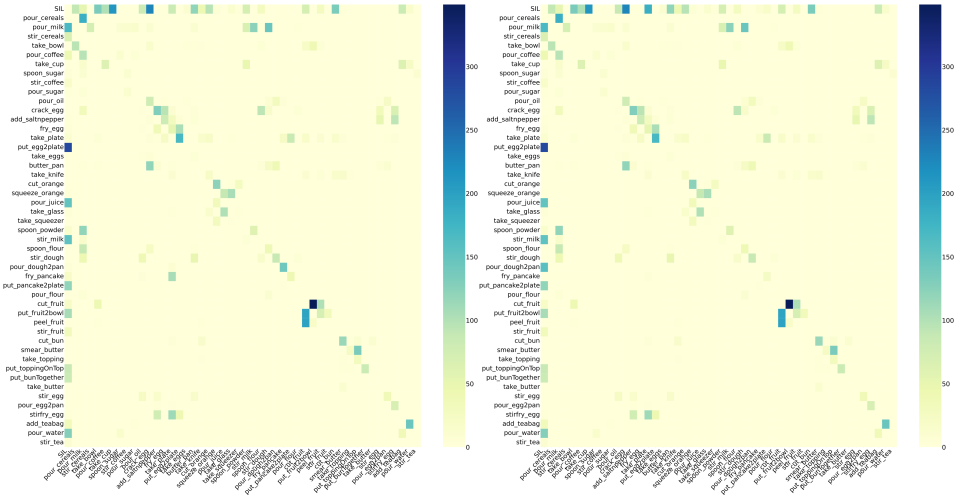

Figure A1: **Heatmap of the frequency of 2-gram action label pairs in Breakfast.** The left is the original dataset and the right is the dataset with the action masking technique. We use initial action as'pour_dough2pan.' The former action is represented on the Y-axis and the latter action on the X-axis.

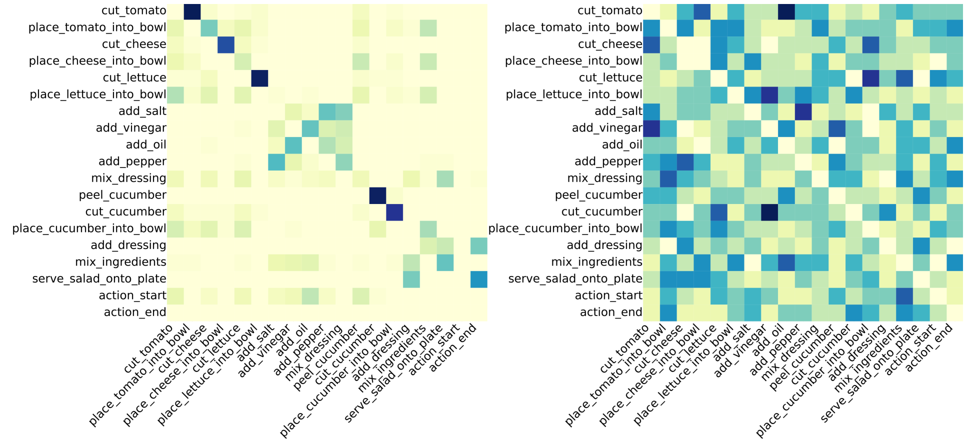

Figure A2: **Heatmap of the frequency of 2-gram action label pairs in the 50Salads dataset.** The left displays the original dataset and the right shows the dataset with the Shuffle technique. The former action is represented on the Y-axis and the latter action on the X-axis.

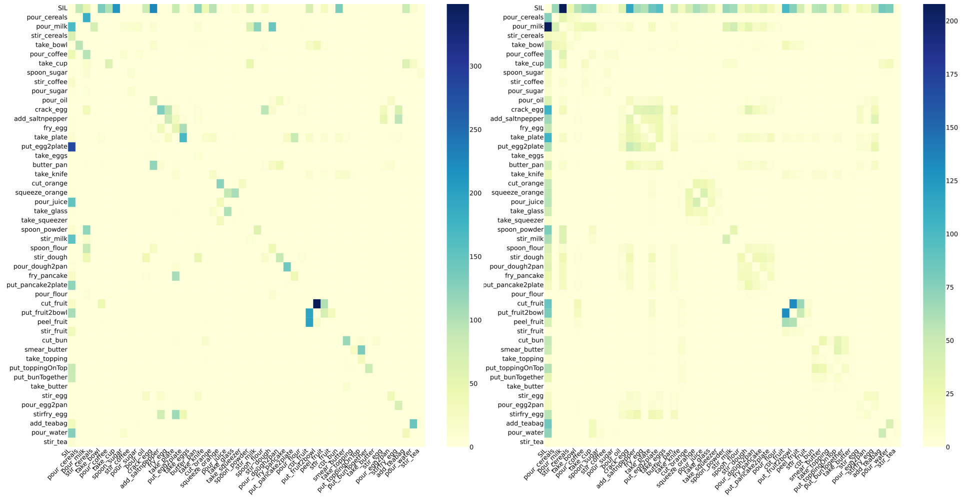

Figure A3: **Heatmap of the frequency of 2-gram action label pairs in the Breakfast dataset.** The left displays the original dataset and the right represents the dataset with the Shuffle technique. The former action is represented on the Y-axis and the latter action on the X-axis.

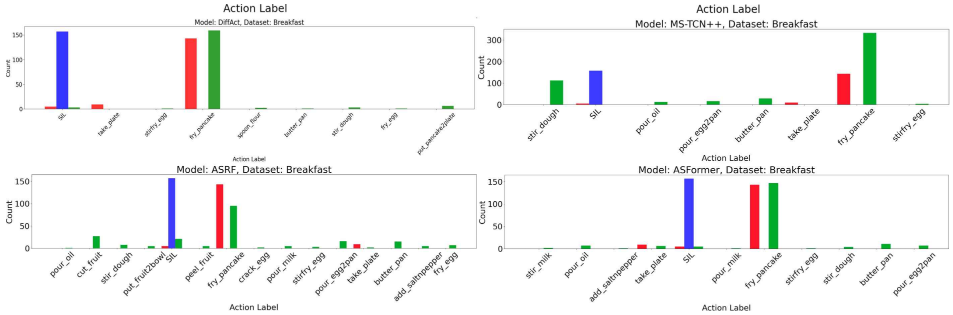

Figure A4: **Distribution of predicted action labels with four models on Breakfast dataset** The red bar represents the count in the original video set; the blue bar displays the count of ground truth label in the masked video set used for evaluation, where the latter action label is replaced with 'no-action' ('SIL' in Breakfast). The green bar represents the count of the model's prediction for the masked video section.

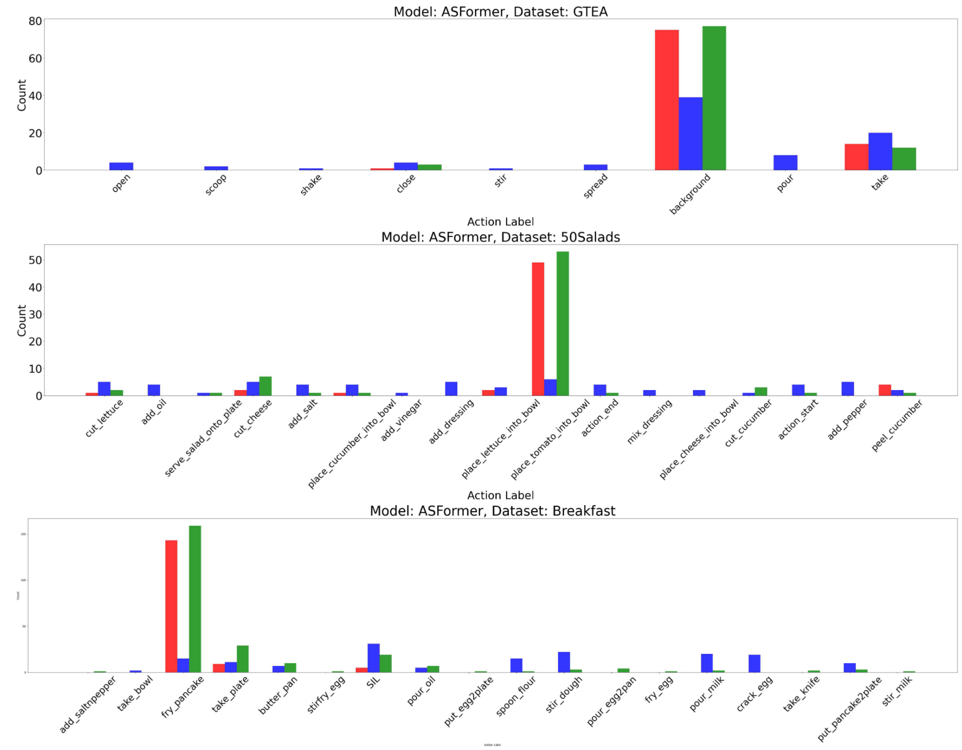

Figure A5: **Distribution of predicted action labels on ASFormer with various datasets.**

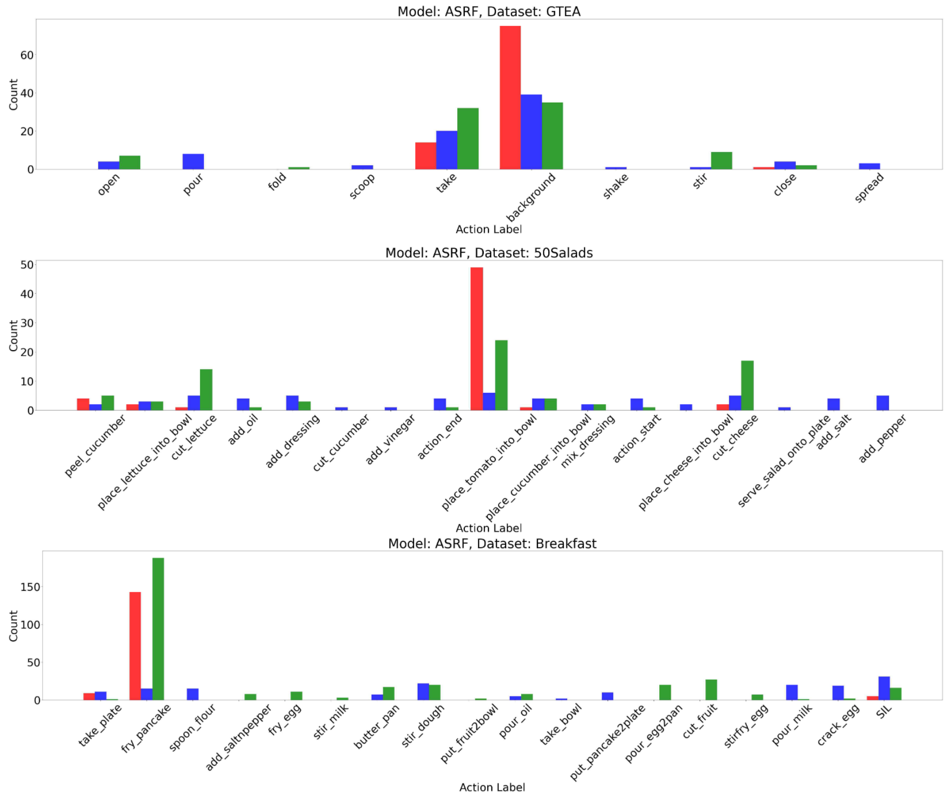

Figure A6: **Distribution of predicted action labels on ASRF with various datasets.**

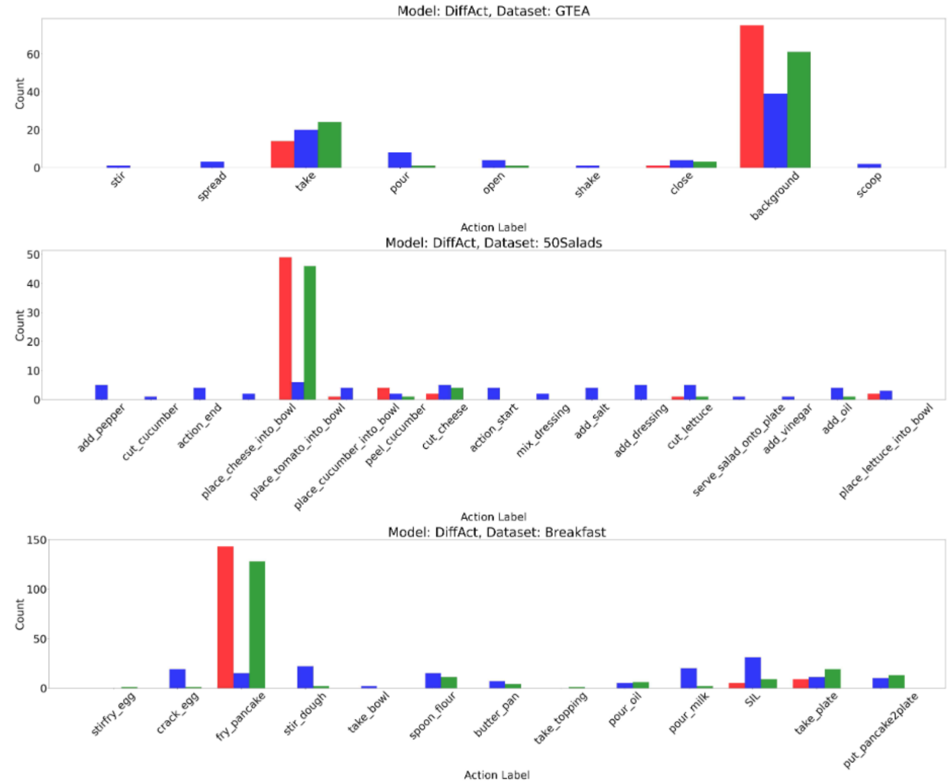

Figure A7: **Distribution of predicted action labels on DiffAct with various datasets.**

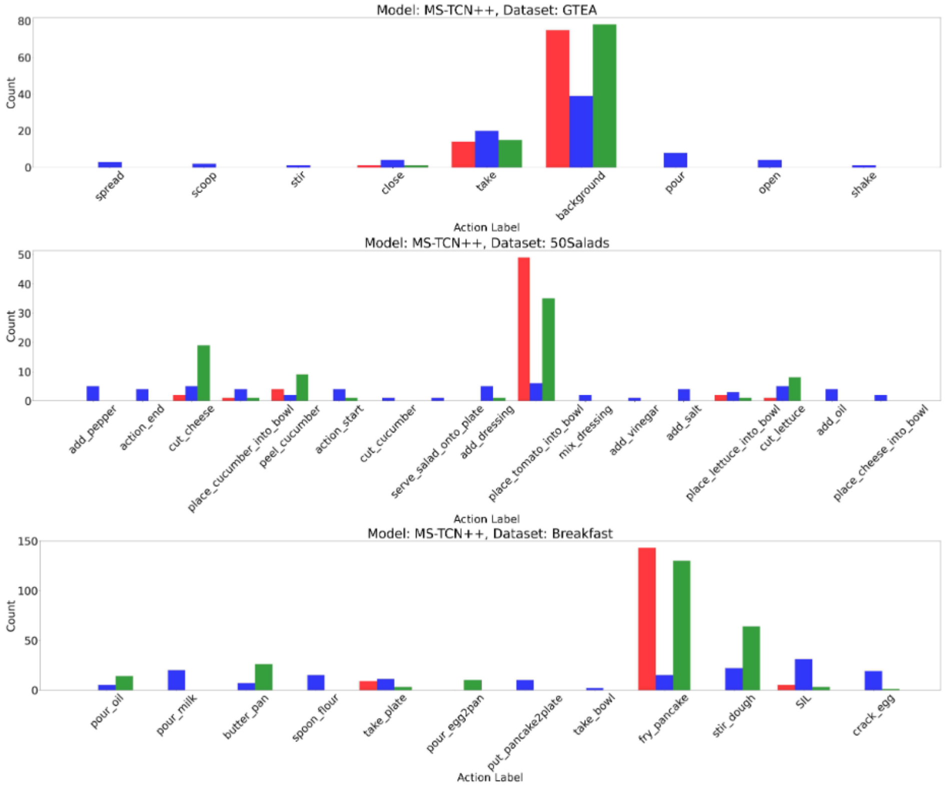

Figure A8: **Distribution of predicted action labels on MS-TCN++ with various datasets.**