Exploring Ordinal Bias in Action Recognition for Instructional Videos

Joochan Kim[1], Minjoon Jung[2] & Byoung-Tak Zhang[2]

[1] Korea Institute of Science and Technology, [2] Seoul National University

Poster Presenter: Joochan Kim (joochan.k@kist.re.kr)

ICLR

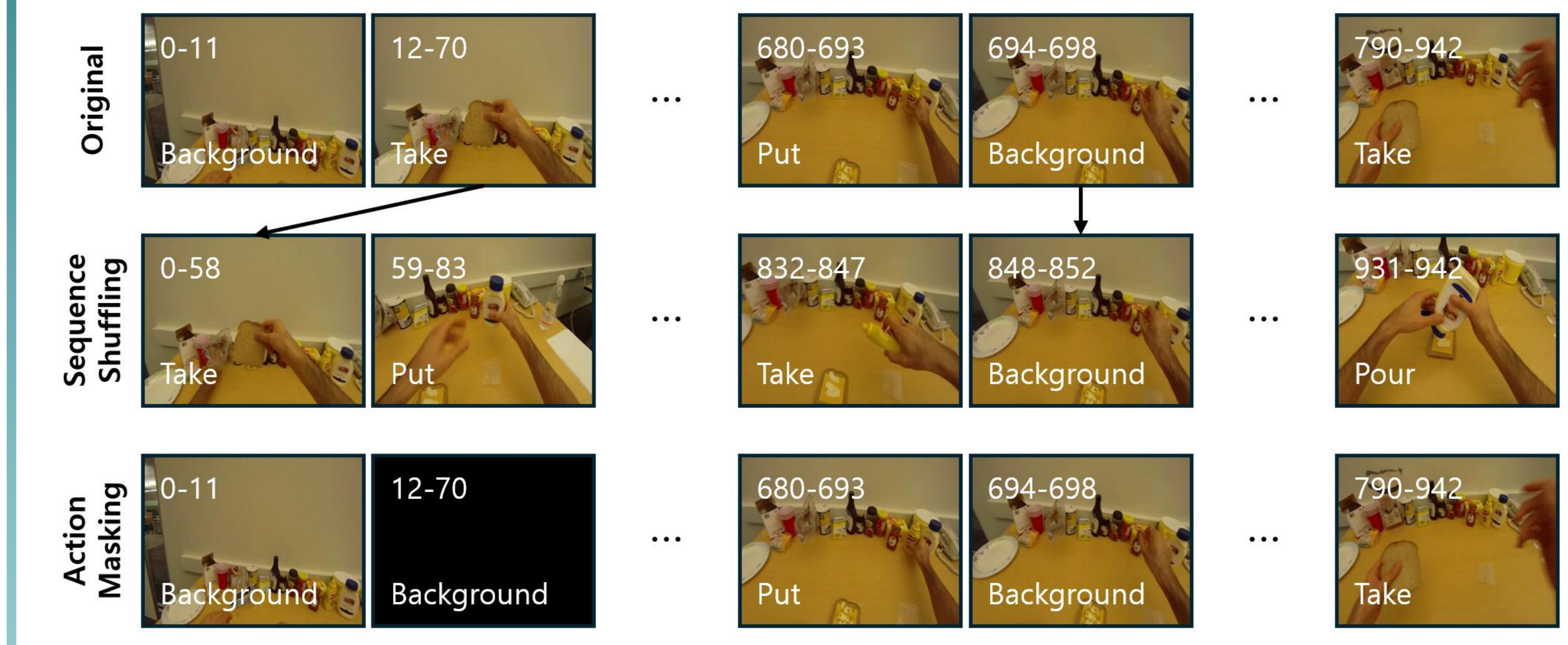
Abstract

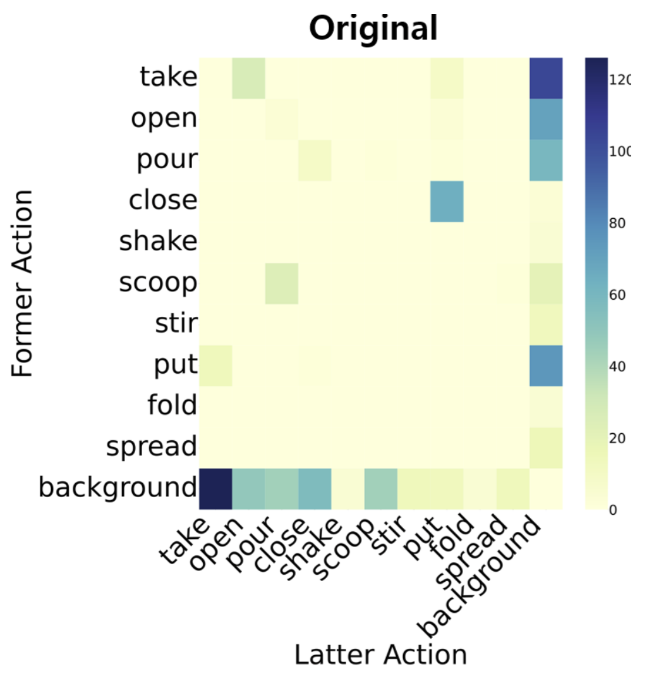
- We define *ordinal bias* as a phenomenon where instructional video action recognition models rely on dominant action sequence patterns rather than true video comprehension, and propose *Action Masking* and *Sequence Shuffling* as systematic evaluation methods.
- Our experiments demonstrate significant performance drops when models face non-standard action sequences, highlighting vulnerability to ordinal bias and emphasizing the need for more robust evaluation frameworks and models.

Introduction

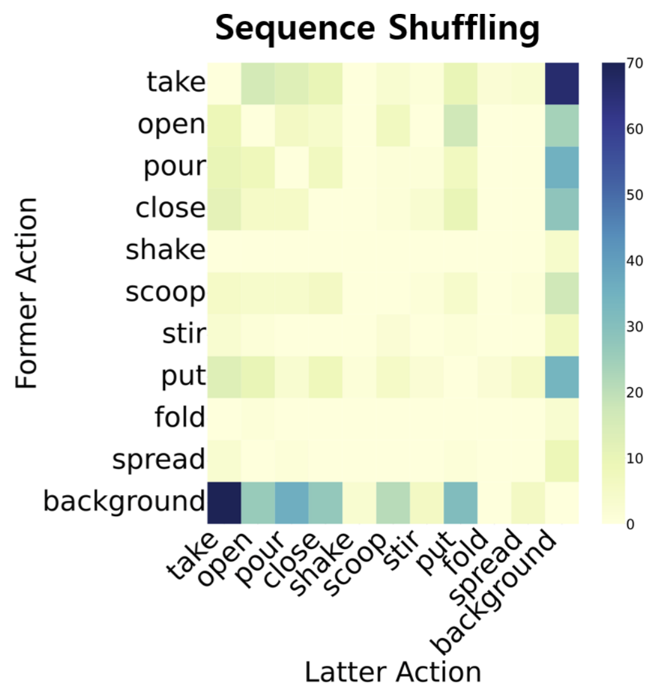
- Action recognition models may rely on repetitive dataset patterns rather than actual video understanding, potentially overestimating performance in understanding publicly released instructional video datasets.
- Analysis reveals highly skewed action pair distributions, causing models to predict based on learned sequence patterns rather than visual cues.
- We propose two video manipulation techniques to measure ordinal bias dependency, revealing model weaknesses and highlighting the need for developing models capable of generalizing beyond fixed action patterns.

Video Manipulation Techniques

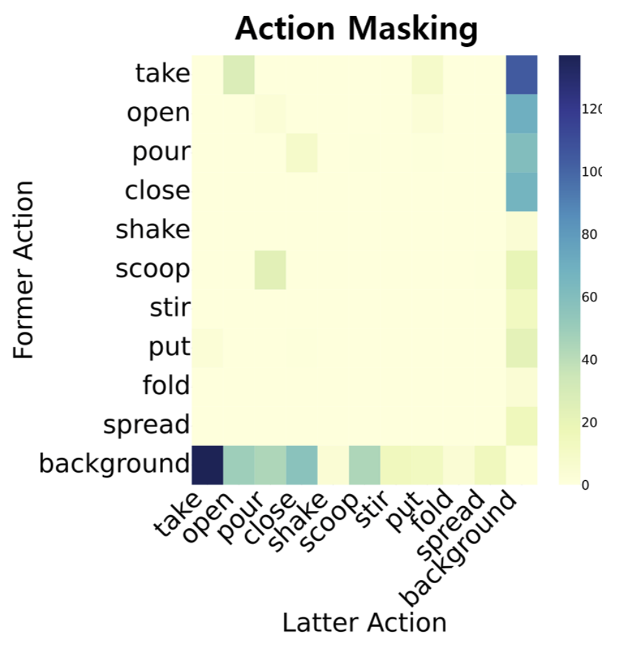




- **Original**
  - Instructional videos that demonstrate step-by-step instructions for completing various tasks such as cooking (e.g., GTEA[1]).
- **Sequence Shuffling**
  - Randomly rearrange the order of action segments while preserving frame order within each action unit.
  - Maintains internal semantic coherence while challenging models' dependency on fixed sequence patterns.

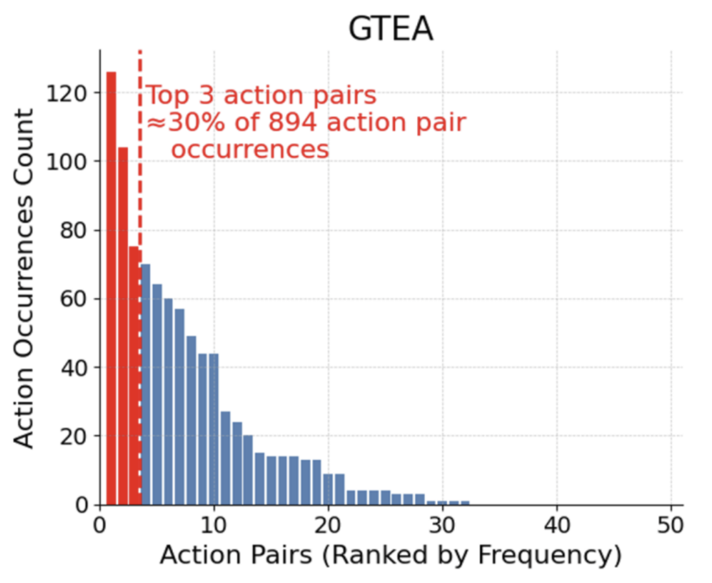


- **Action Masking**
  - Masks video frames of a specific action unit and replace the corresponding action label with 'no action.'
  - Compels models to rely on alternative visual cues rather than dominant action sequence patterns.

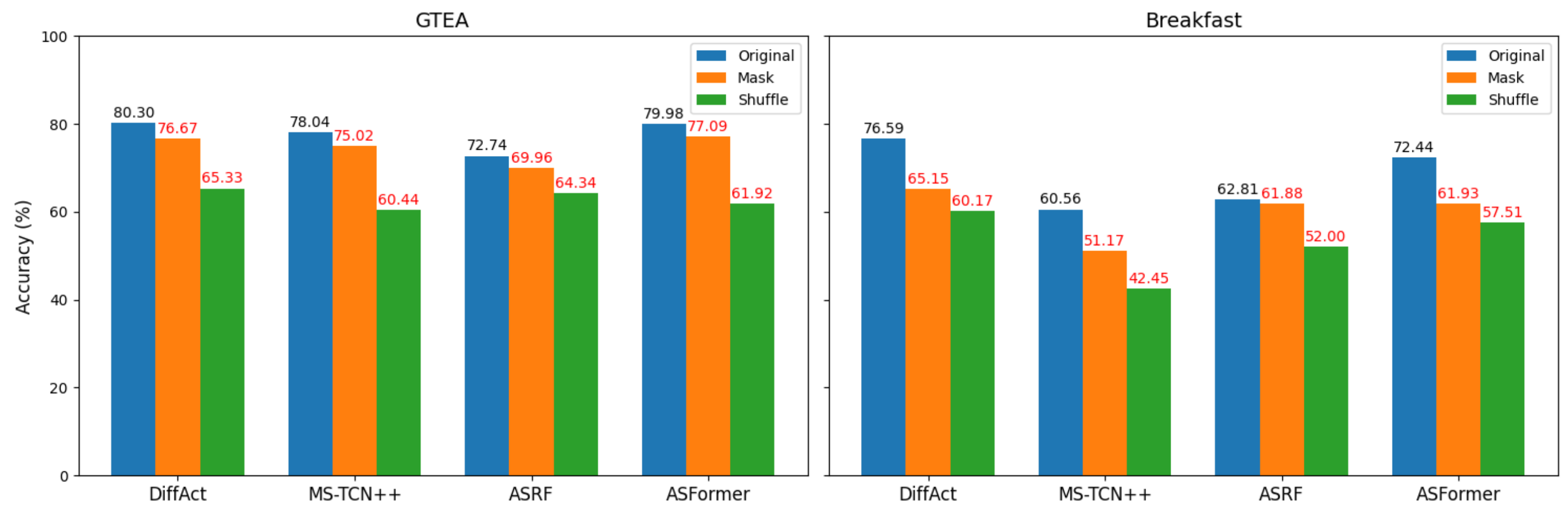


Experiments

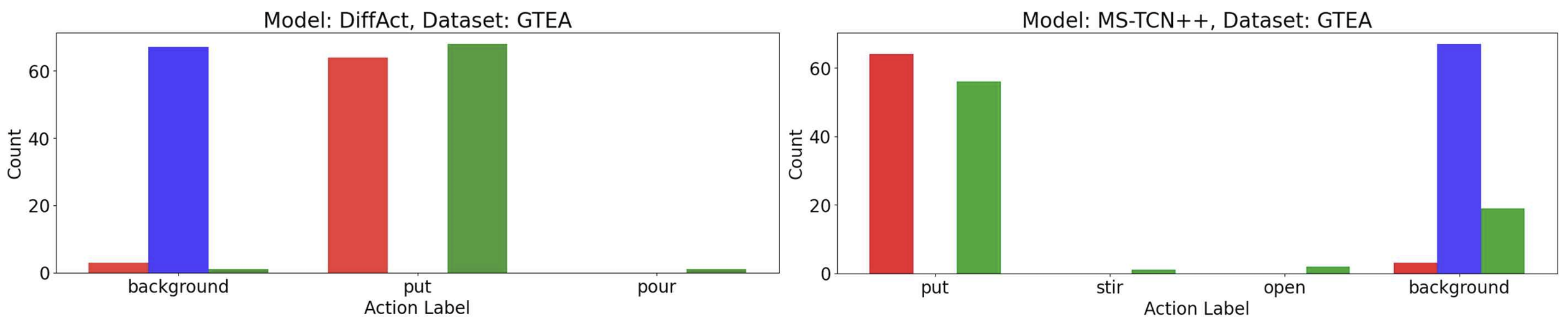
- **Distribution of Dataset**
  - Analyzed three instructional video datasets and found that only 3 of 32 action pairs are dominating 30% of the all occurrences, indicating *long-tailed* distribution.



- **Quantitative Results**
  - Models like ASFormer[2] experience performance degradation on the proposed manipulation set, showing biased correlation.



- **Qualitative Results**
  - Models tend to make predictions (Green) with trends in the training data (Red), rather than with given visual cues (Blue).



- **Additional Training with Augmented Dataset**
  - Failed to generalize well on given manipulated dataset, indicating that further training is not a definite solution.

Dataset	MS-TCN++				ASFormer			
	O/O	C/O	C/S	C/M	O/O	C/O	C/S	C/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	-	-	-

Conclusion

- **Key Findings**
  - Current action recognition models rely on dataset-specific action sequences rather than true video understanding.
  - Models show significant performance drops when faced with non-standard action sequences.
  - Data augmentation alone fails to mitigate this bias, indicating deeper architectural issues.
- **Implications & Future Work**
  - Benchmark accuracy metrics likely overestimate real-world performance.
  - Development of models that can generalize beyond fixed action patterns is needed.
  - Construction of a more automatic approach to identify the existence of ordinal bias would be beneficial.

References

[1] Fathi, Alireza, Xiaofeng Ren, and James M. Rehg. "Learning to recognize objects in egocentric activities." CVPR 2011. IEEE, 2011.

[2] Yi, Fangqiu, Hongyu Wen, and Tingting Jiang. "ASFormer: Transformer for Action Segmentation." (2021).