Supplementary Material

An Empirical Study on Factors of Influence for Single-Frame Supervised Temporal Action Detection

1 Qualitative Comparisons on GTEA [5], BEOID [2], ActivityNet1.2 [1], ActivityNet1.3 [1], Multi-THUMOS [8], 50 Salads [7], MPII Cooking 2 [6], and Breakfast [3] Datasets

Figure 1, 2, 3, 4, 5, 6, 7, and 8 show qualitative comparisons with LACP [4] on GTEA [5], BEOID [2], ActivityNet1.2 [1], ActivityNet1.3 [1], Multi-THUMOS [8], 50 Salads [7], MPII Cooking 2 [6], and Breakfast [3] datasets, respectively. With respect to qualitative results, our SF-baseline outperforms the previous best state-of-the-art approach [4].

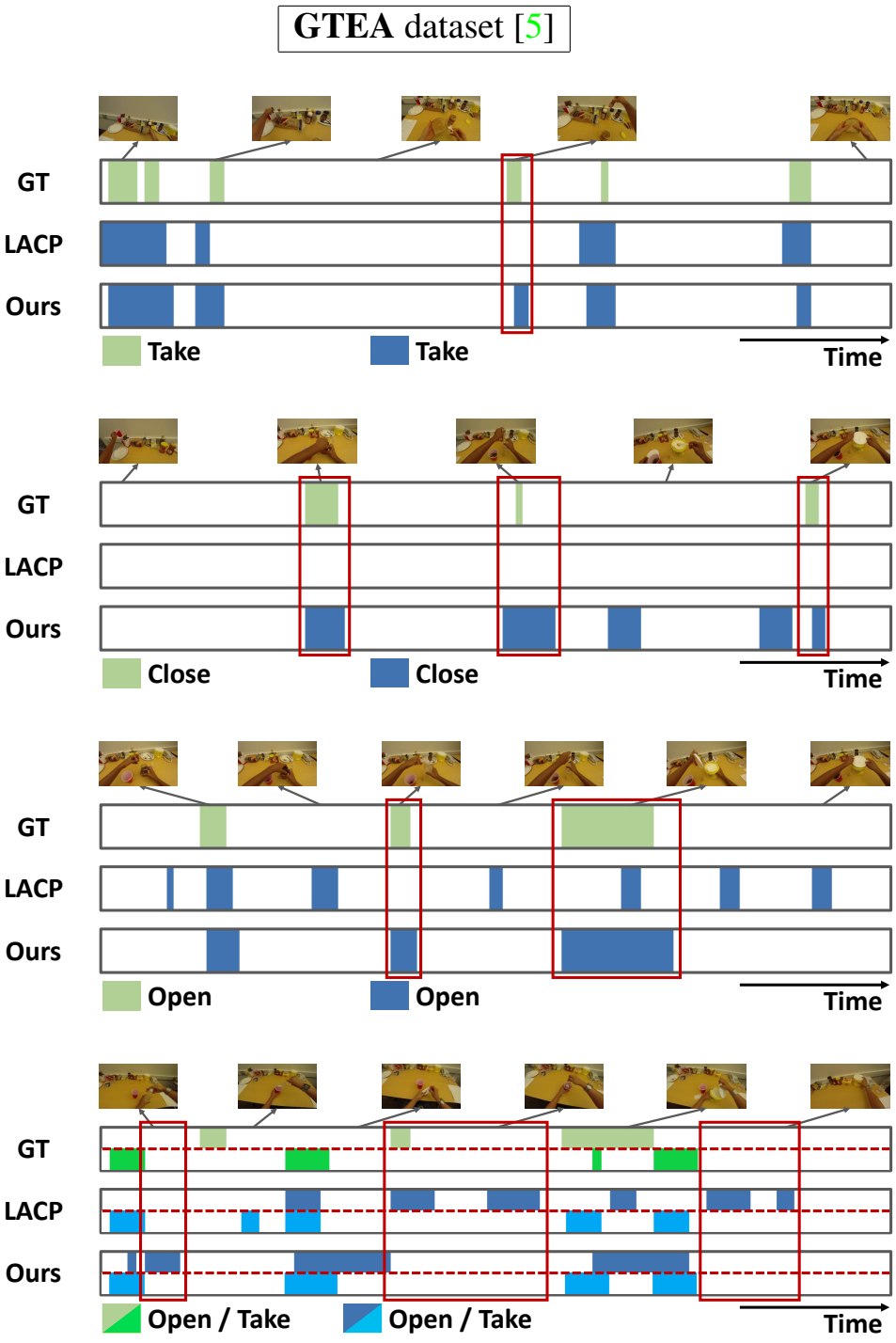

Figure 1: Qualitative comparison with LACP [4] on **GTEA** dataset [5].

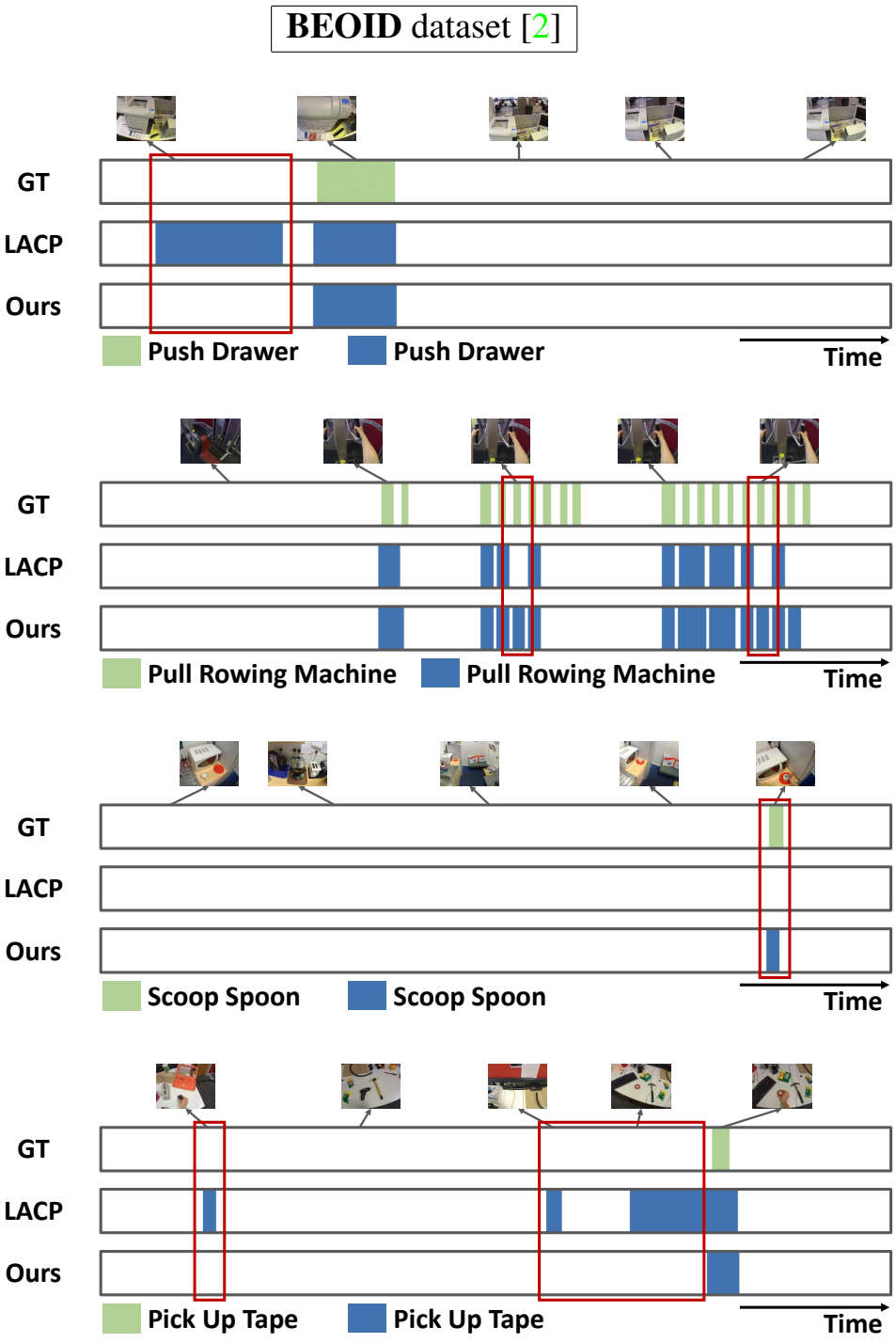

Figure 2: Qualitative comparison with LACP [4] on **BEOID** dataset [2].

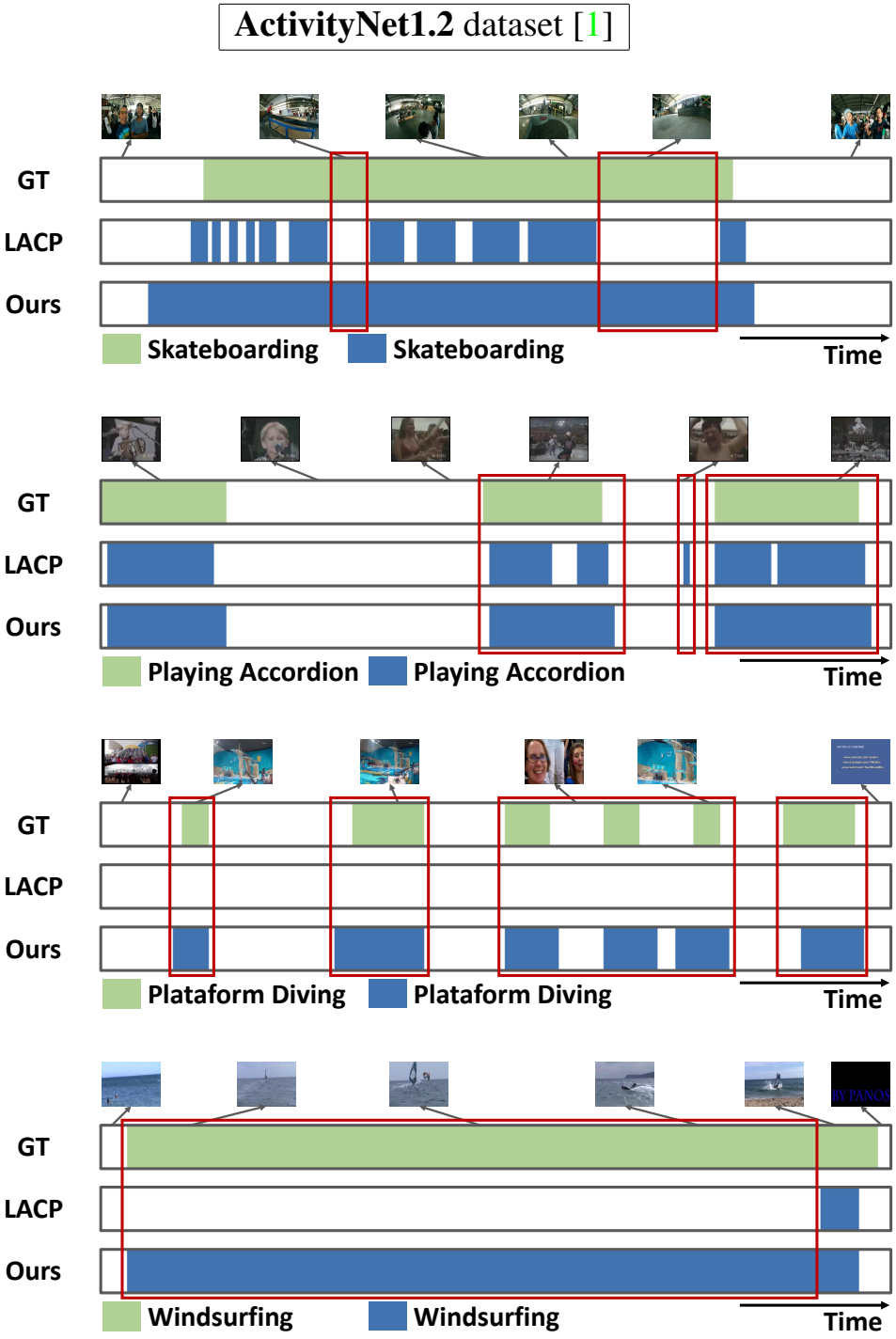

Figure 3: Qualitative comparison with LACP [4] on **ActivityNet1.2** dataset [1].

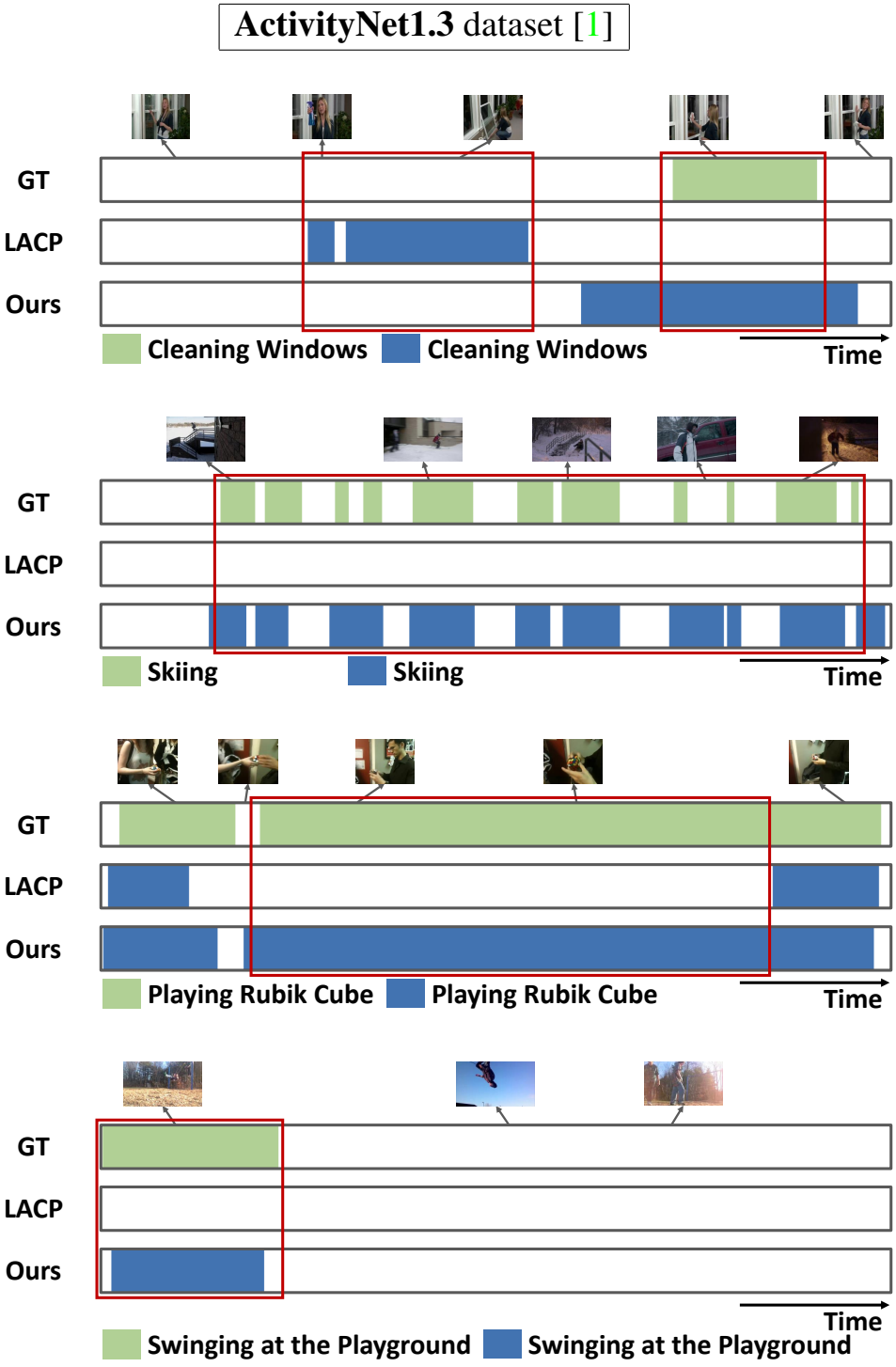

Figure 4: Qualitative comparison with LACP [4] on **ActivityNet1.3** dataset [1].

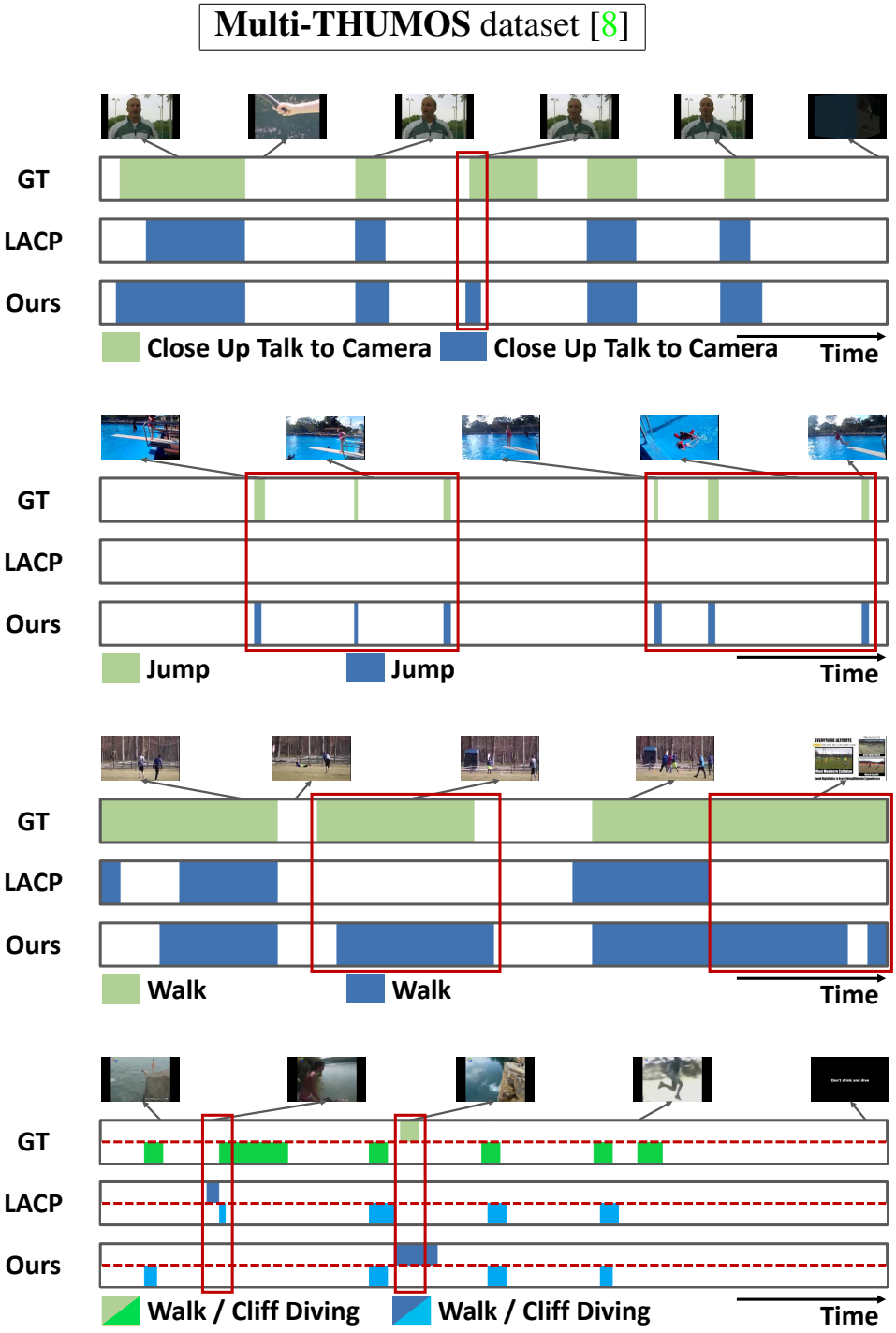

Figure 5: Qualitative comparison with LACP [4] on **Multi-THUMOS** dataset [8].

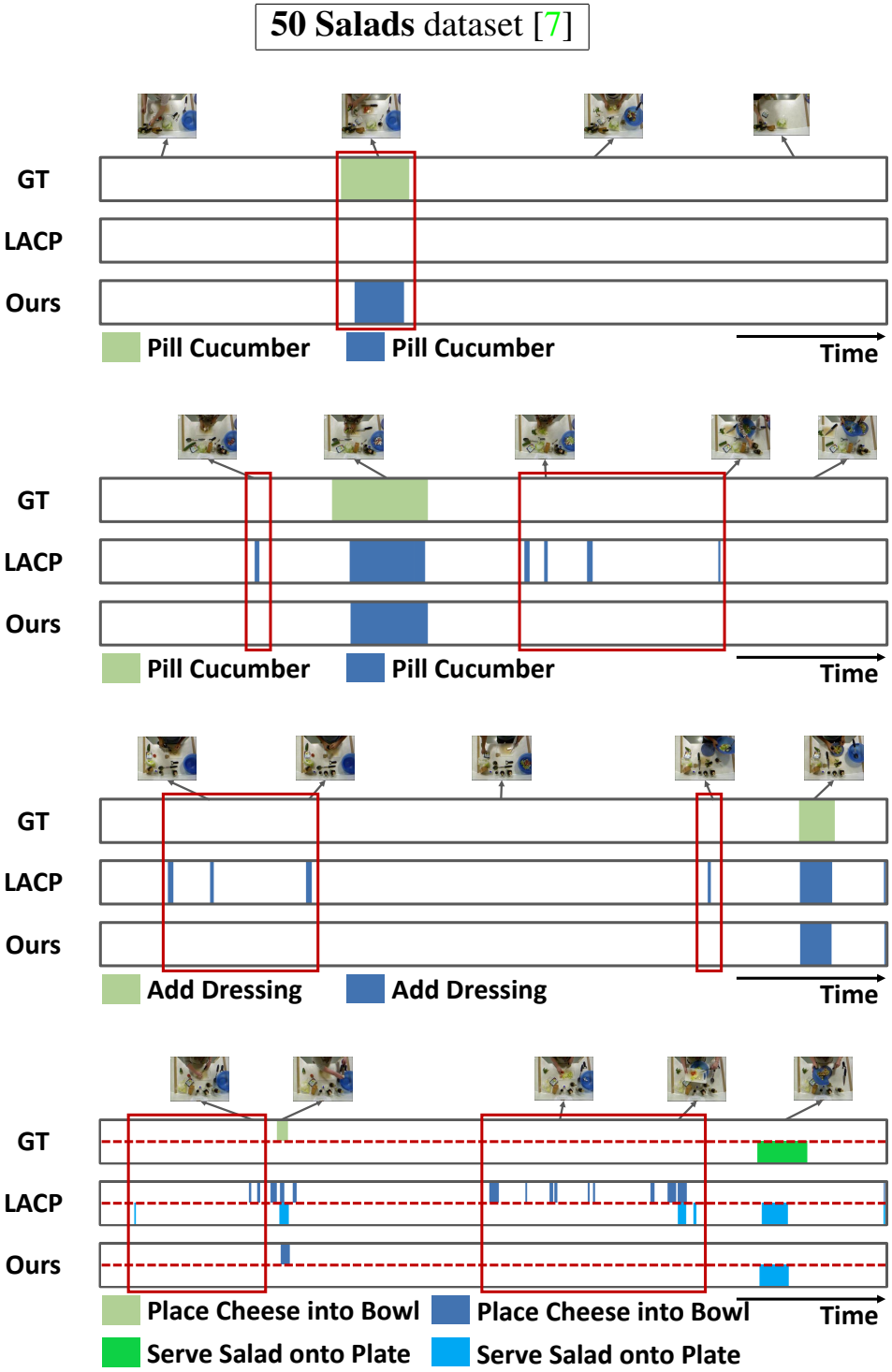

Figure 6: Qualitative comparison with LACP [4] on **50 Salads** dataset [7].

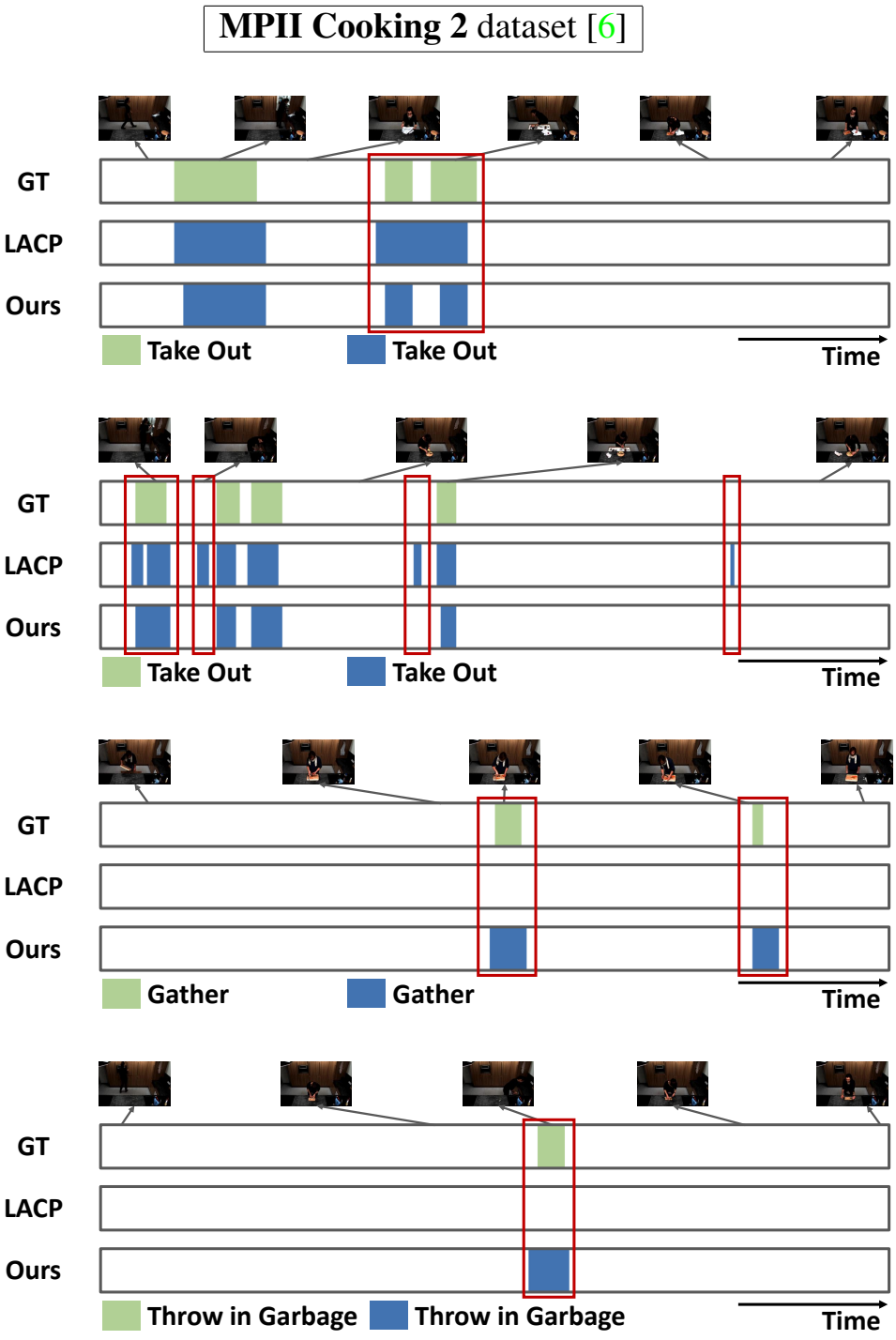

Figure 7: Qualitative comparison with LACP [4] on **MPII Cooking 2** dataset [6].

Figure 8: Qualitative comparison with LACP [4] on **Breakfast** dataset [3].

References

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *CVPR*, 2015. 2, 5, 6

[2] Dima Damen, Teesid Leelasawassuk, Osian Haines, Andrew Calway, and Walterio W Mayol-Cuevas. You-Do, I-Learn: Discovering Task Relevant Objects and their Modes of Interaction from Multi-User Egocentric Video. In *BMVC*, 2014. 2, 4

[3] Hilde Kuehne, Ali Arslan, and Thomas Serre. The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities. In *CVPR*, 2014. 2, 10

[4] Pilhyeon Lee and Hyeran Byun. Learning Action Completeness from Points for Weakly-supervised Temporal Action Localization. In *ICCV*, 2021. 2, 3, 4, 5, 6, 7, 8, 9, 10

[5] Peng Lei and Sinisa Todorovic. Temporal Deformable Residual Networks for Action Segmentation in Videos. In *CVPR*, 2018. 2, 3

[6] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. International Journal of Computer Vision, 119:346–373, 2016. 2, 9

[7] Sebastian Stein and Stephen J McKenna. Combining Embedded Accelerometers with Computer Vision for Recognizing Food Preparation Activities. In *CVPR*, 2013. 2, 8

[8] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every Moment Counts: Dense Detailed Labeling of Actions in Complex Videos. International Journal of Computer Vision, 126(2):375–389, 2018. 2, 7