Supplementary Material
DEFT-DPT: A Lightweight Multimodal Framework for Egocentric Video Action Recognition

Pawanesh Kumar Vishwakarma, Dushyant Kumar Singh, Senior Member, IEEE, Abhimanyu Sahu*

Department of Computer Science & Engineering, Motilal Nehru National Institute of Technology Allahabad, Prayagraj, India-211004

(e-mails: pawanesh.2023rcs04@mnnit.ac.in, dushyant@mnnit.ac.in, abhimanyus@mnnit.ac.in)

In this supplementary material, we provide detailed description of the datasets and additional qualitative results; those are omitted in the main paper.

Section A: Detailed Information on the Experimented Datasets

- Detailed information about the EPIC-Kitchens dataset
- Detailed information about the Meccano dataset
- Detailed information about the GTEA dataset
- Detailed information about the ADL dataset

Section B: Additional Qualitative Results and Ablation Studies for Action Recognition

- Qualitative results from the first ablation study on the Meccano dataset
- Qualitative results from the first ablation study on the GTEA dataset
- Qualitative results from the first ablation study on the ADL dataset

Section C: Additional qualitative results for Action Localization

- Qualitative localization results on the GTEA dataset
- Qualitative localization results on the ADL dataset

Section D: Additional Qualitative Results for Failure Case Analysis

- Qualitative results on the EPIC-kitchen dataset

A. Detailed Information on the Experimented Datasets

• **EPIC-Kitchens [1]** The EPIC-Kitchens dataset consists of egocentric videos capturing daily kitchen activities performed by different individuals in their home environments. It contains 432 videos recorded by 32 participants, each performing unscripted kitchen-related tasks. The videos are captured using a head-mounted GoPro camera with an adjustable mount, allowing control over the viewpoint to adapt to diverse kitchen layouts and lighting conditions. The dataset is annotated for 352 object categories and 125 action classes, providing a rich foundation for studying complex hand–object interactions and multimodal learning in egocentric scenarios. In total, it includes more than 28K annotated video segments covering a wide variety of cooking and interaction sequences. All videos are provided in MPEG-1 format with a spatial resolution of 1920×1080 at 59.94 fps. Owing to its scale, diversity, and naturalistic recordings, EPIC-Kitchens serves as one of the largest and most comprehensive benchmarks for egocentric action recognition and temporal activity understanding.

• **Meccano [2]** The Meccano dataset focuses on industrial assembly tasks, featuring 20 participants assembling a toy model using a head-mounted camera. It includes both RGB and depth modalities, enabling multimodal learning and cross-view understanding. The dataset contains more than 8,000 annotated action instances covering 14 distinct action classes. The Meccano dataset is particularly valuable for studying hand–object interactions, robotic skill learning, and task-oriented visual reasoning. Its real-world complexity introduces several challenges, including fine-grained action segmentation, frequent hand occlusions, and varying lighting conditions due to dynamic environments.

• **GTEA [3]** The GTEA dataset contains 17 video sequences recorded from 14 different subjects performing meal preparation tasks. Each sequence lasts approximately four minutes and is recorded using Tobii eye-tracking glasses equipped with an outward-facing camera. The videos are captured at a frame rate of 30 fps with a spatial resolution of 480×640. The dataset focuses exclusively on meal preparation activities, emphasizing hand–object interactions that are crucial for egocentric activity recognition. A total of 30 distinct food items and kitchen objects are involved across all recordings, making it suitable for studying task-specific visual attention and object manipulation.

• **ADL [4]** The ADL (Activities of Daily Living) dataset comprises 20 videos covering 10 distinct daily activities performed by 20 individuals in a laboratory kitchen setting. Videos are recorded using a chest-mounted camera positioned to face the working counter. Each sequence includes detailed annotations for both activity and object categories.

Pirsiavash and Ramanan [4] introduced this dataset to benchmark algorithms for joint action and object recognition. All videos are provided in MPEG-1 format with a resolution of 960×1280 at 30 fps, and a wide 170° field of view. The ADL dataset is widely used for modeling egocentric interactions involving routine household activities under controlled but realistic conditions.

B. Additional Qualitative Results and Ablation Studies for Action Recognition

Meccano first ablation study

The qualitative results from the first ablation study on the EPIC-Kitchens dataset, shown in Fig. 1, show that our DEFT-based model accurately recognizes actions such as *"take_wrench"*, *"take_bolt"*, and *"align_object"*, while the baseline often misclassifies them as visually similar actions like *"take_handlebar"*, *"plug_screw"*, or *"drink_water"*. The confusion matrix in Fig. 2 further illustrates that our approach reduces inter-class confusion, yielding a clearer and more diagonalized structure that highlights the robustness of DEFT in enhancing action discrimination.

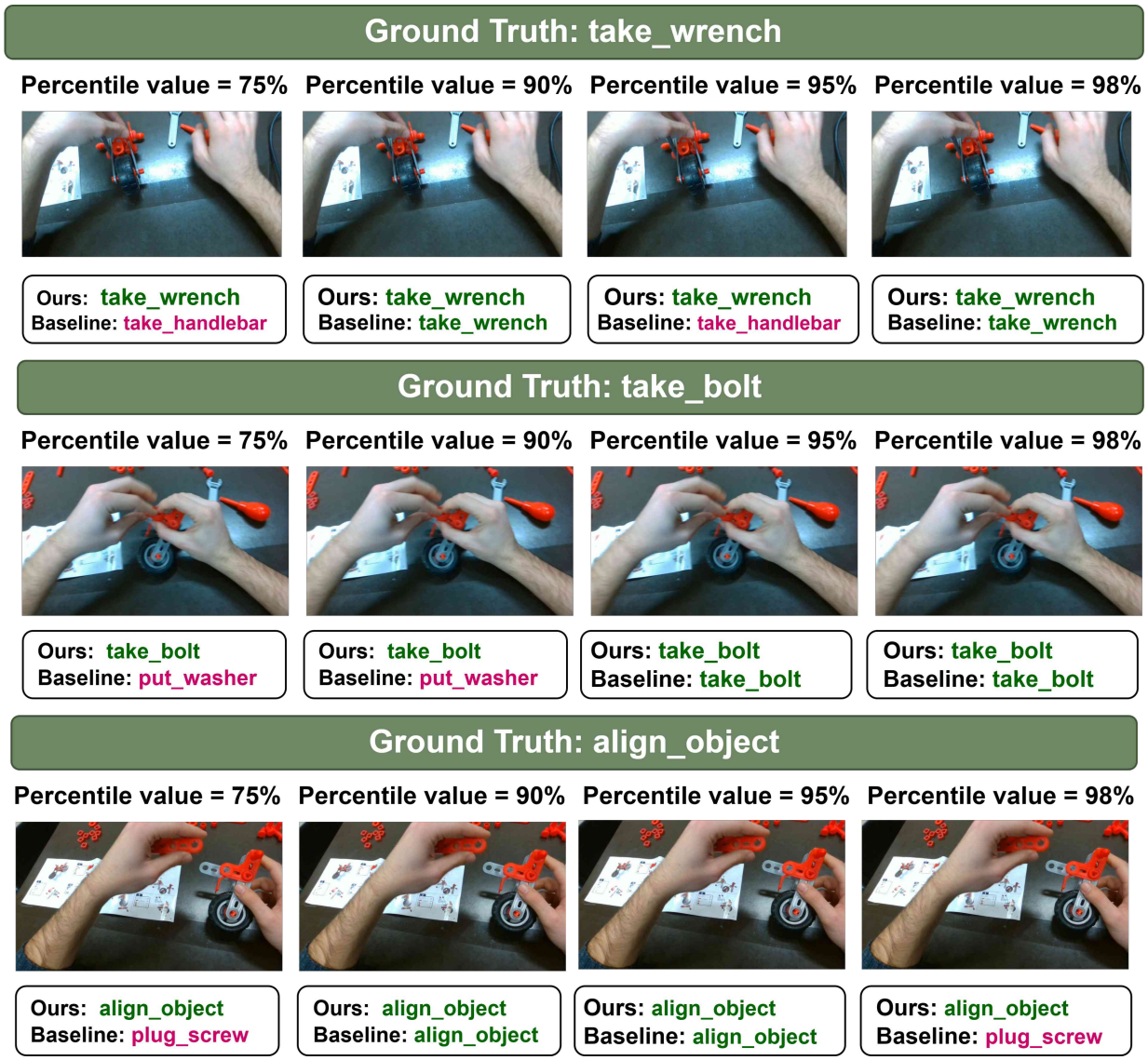

Figure 1: Action predictions on video 0005 from Meccano dataset [2] of our model with and without DEFT (denoted as the baseline) over percentile value with three testing examples

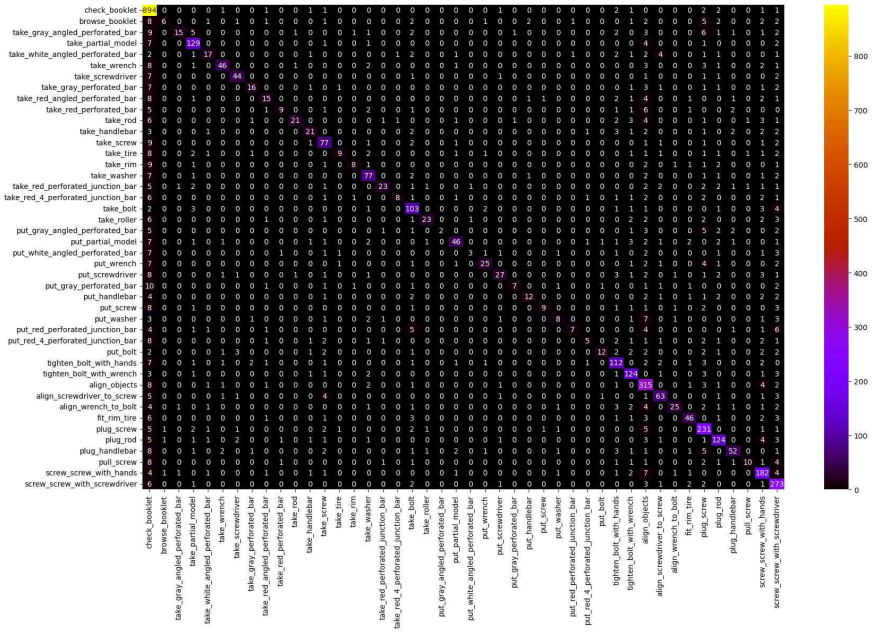

Figure 2: Confusion matrix of our model for all 44 classes on 0005 video from Meccano [2] dataset.

GTEA first ablation study

The qualitative results from the first ablation study on the GTEA dataset, shown in Fig. 3, indicate that our DEFT-based model accurately recognizes actions such as *"open_mustard"*, *"take_ketchup"*, and *"take_hotdog"*, while the baseline often misclassifies them as visually similar actions like *"put_ketchup"* or *"put_bread"*. The confusion matrix in Fig. 4 further demonstrates that our approach minimizes inter-class confusion, producing a clearer and more diagonalized structure that highlights the robustness of DEFT in enhancing action discrimination.

ADL first ablation study

The qualitative results from the first ablation study on the ADL dataset, shown in Fig. 5, show that our DEFT-based model accurately recognizes actions such as *"making_tea"* and *"using_computer"*, whereas the baseline often misclassifies them as visually similar actions like *"drinking_coffee"* or *"using_cell"*. The confusion matrix in Fig. 6 further demonstrates that our approach reduces inter-class confusion, yielding a clearer and more diagonalized structure that confirms the robustness of DEFT in enhancing action discrimination.

C. Additional Qualitative Results for Action Localization

A key strength of our DEFT-based framework is its ability to spatially localize actions in egocentric video frames. In Fig. 7, DEFT accurately focuses on the relevant hand–object interaction regions for actions such as *"open mustard"* (label 9), *"take ketchup"* (label 49) and *"take hotdog"* (label 47) in the GTEA dataset.

Figure 3: Action predictions on video S1_HotDog_C1 from GTEA dataset [3] of our model with and without DEFT (denoted as the baseline) over percentile value with three testing examples

A key strength of our DEFT-based framework is its ability to spatially localize actions in egocentric video frames. In Fig. 8, DEFT accurately focuses on the relevant hand–object interaction regions for actions such as *"making tea"* (label 12), *"using computer"* (label 23) and *"washing hand"* (label 5) in the ADL dataset.

D. Additional Qualitative Results for Failure Case Analysis

We present two representative failure cases in Fig. 9. In the first row, the model misclassifies *"take-up_liquid"* as *"take_bottle"* due to significant occlusion and similar hand–object motions, while in the second row, *"pull-down_tablecloth"* is incorrectly predicted as *"close_cupboard"* because of motion blur and visual clutter. These cases highlight common challenges in egocentric video un-

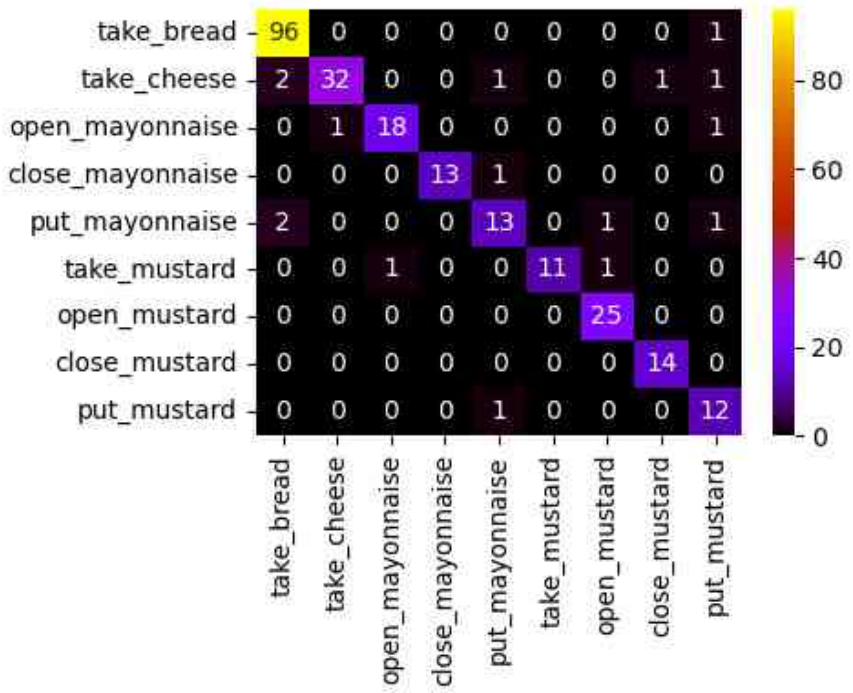

Figure 4: Confusion matrix of our model for all 09 classes on S2_cheese_C1 video from GTEA [3] dataset.

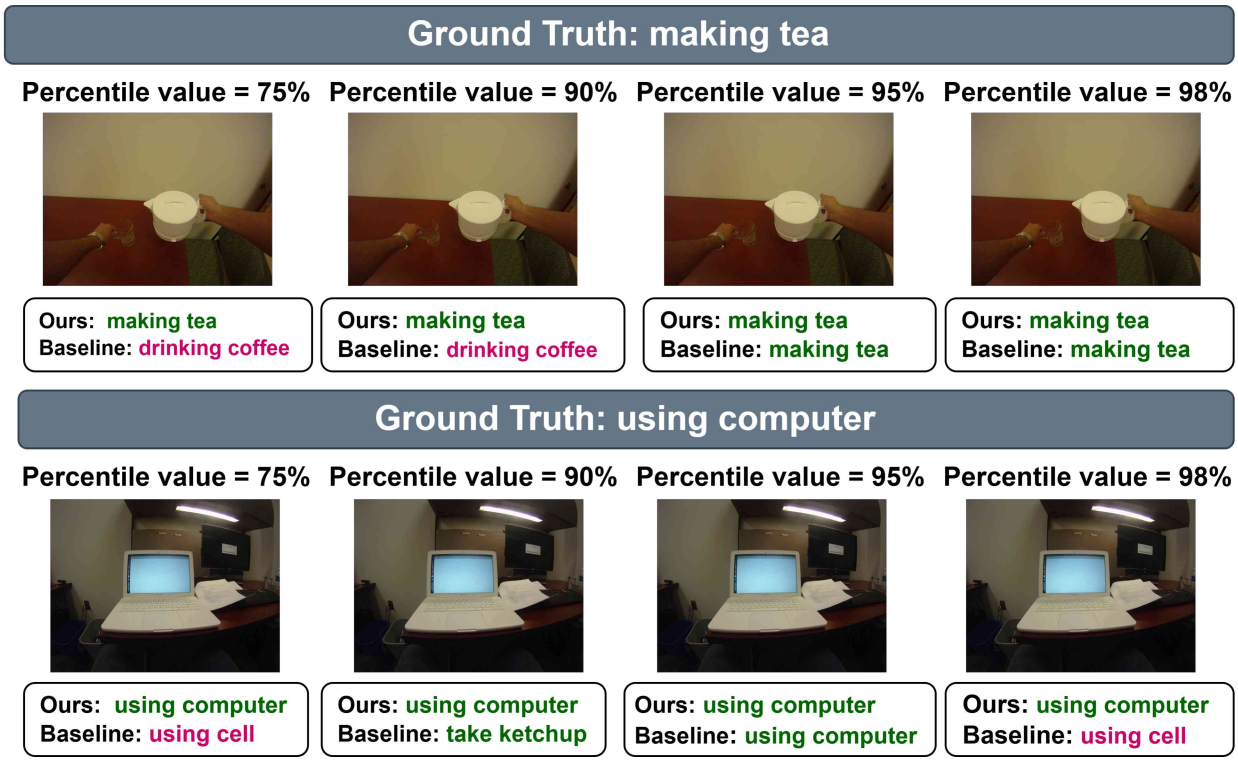

Figure 5: Action predictions on video P_11 from ADL dataset [4] of our model with and without DEFT (denoted as the baseline) over percentile value with three testing examples

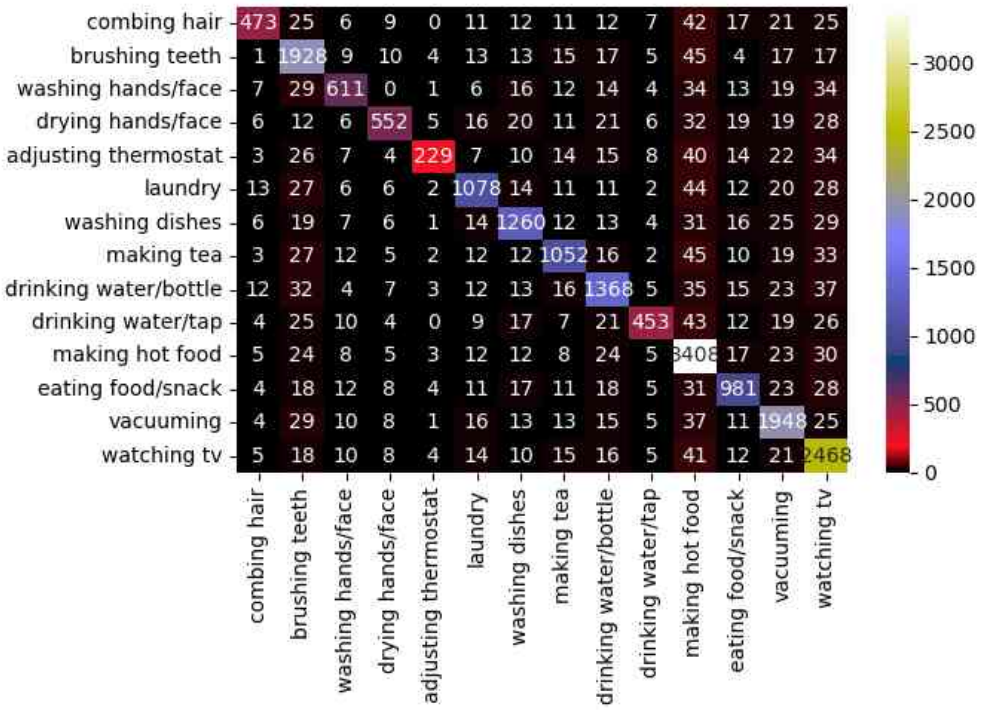

Figure 6: Confusion matrix of our model for all 14 classes on P_16 video from ADL dataset [4].

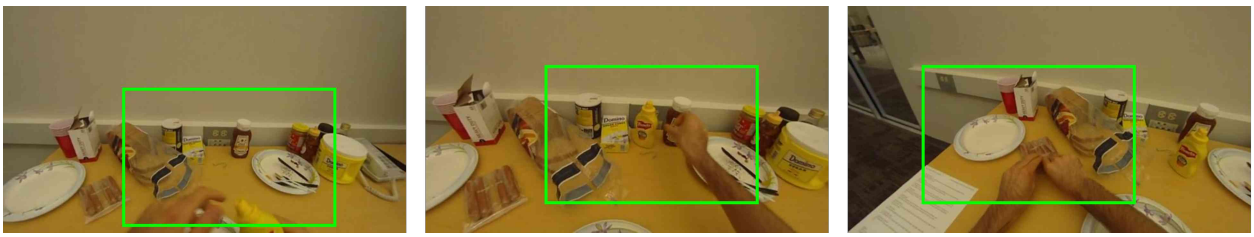

Figure 7: Localization of simple actions in S1_HotDog_C1 video of the GTEA dataset: *"open mustard"* (label 9), *"take ketchup"* (label 49) and *"take hotdog"* (label 47) (from left to right).

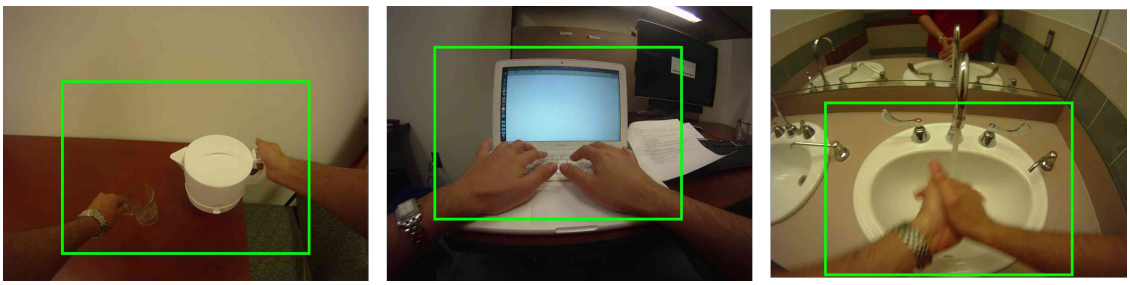

Figure 8: Localization of simple actions in P_11 video of the ADL dataset: *"making tea"* (label 12), *"using computer"* (label 23) and *"washing hand"* (label 5) (from left to right).

derstanding—object ambiguity, occlusion, and transient motion cues—suggesting future directions such as object-centric graph reasoning, motion-aware segmentation, and temporal anticipation to improve model robustness and discrimination.

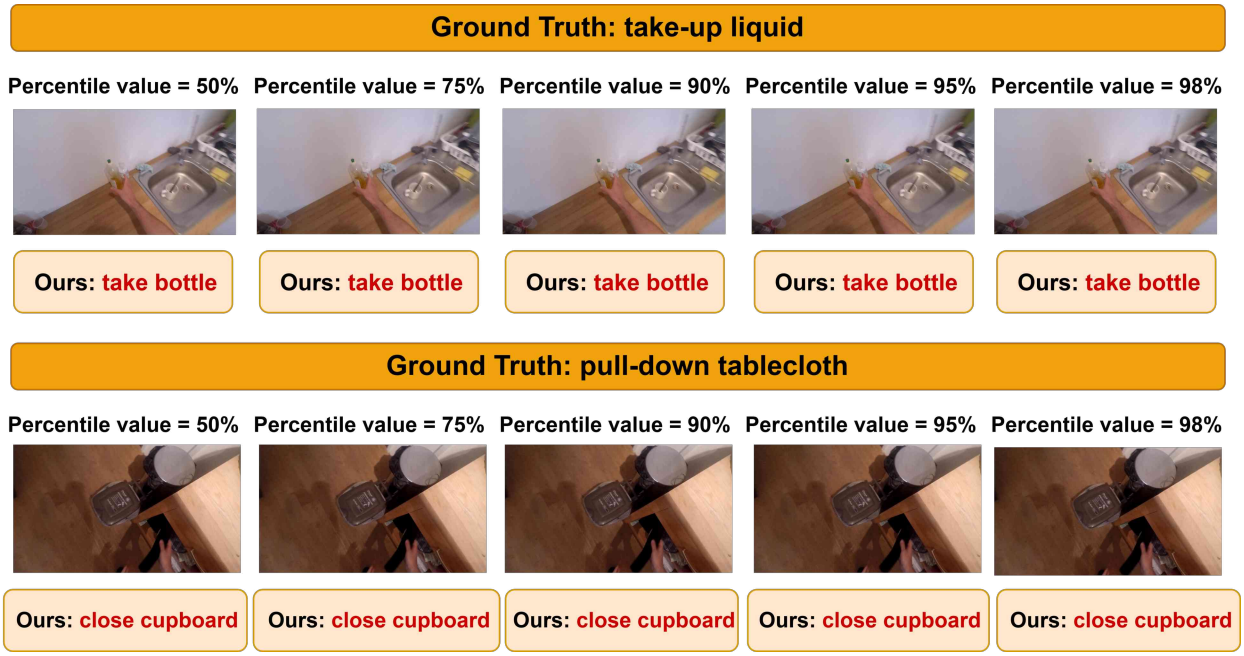

Figure 9: Representative failure cases of the proposed model.

References

[1] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, E. Kazakos, J. Ma, & M. Wray. "Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100", *International Journal of Computer Vision*, 2022, pp. 33-55.

[2] F. Ragusa, A. Furnari, S. Livatino, G. M. & Farinella. "The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain", *in Proc. WACV* 2021, pp. 1569-1578.

[3] Y. Li, Z. Ye, & J. M. Rehg, "Delving into egocentric actions." *in Proc. CVPR*, 2015, pp. 287-295.

[4] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in rst-person camera views", *in Proc. CVPR*, 2012, pp. 2847-2854.