Open Set Video HOI detection from Action-centric Chain-of-Look Prompting
Supplementary Material

1. Introduction

In this supplementary material, we provide:

- More details on the implementation of ACoLP model.
- Examples of qualitative video HOI detection results compared with ground truth.
- Details about the HOI classes related to the interaction class of "*grab*" and "*next_to*".
- Specific HOI splits in open set settings.
- More results of HOI detection in closed set setting.

2. More details on the implementation of ACoLP model.

The adjusting weights μ_1 and μ_2 in Eq.(11) of the main paper are set to be 2.5 and 1, respectively. In open set settings, we split datasets based on the actions contained in HOIs. As shown in Tab. 1, different percentages of actions are utilized for training and testing. For both training set and testing set, the temporal related actions and non-temporal related actions amount for half of the total actions, respectively.

3. Examples of qualitative video HOI detection results compared with ground truth.

We show qualitative examples of video HOI detection results compared with ground-truth in Fig. 1. In each selected video frame, detected human and objects are marked with red bounding boxes and green bounding boxes, respectively. To represent detected HOIs, objects are marked with annotations in the form of "*interaction, object*" in green boxes and person are marked with annotation "*person*" in red boxes. For the sake of clarity, we do not mark the interactions between person in the frames, but list all the detected interactions between person below the frames. The results in Fig. 1 show that our model is able to capture the major *temporal* interactions (e.g. *ride, carry*) in the scene, while neglecting some *spatial* interactions such as *in_front_of*. This is possibly caused by the Dynamic GNN module in our model, which is designed to capture temporal dynamics in videos. However, the Dynamic GNN module is not specifically designed for spatial dynamics, thus spatial interaction results are not as good as temporal interactions.

4. Details about the HOI classes related to the interaction class of "*grab*" and "*next_to*".

In Tab. 3, we list all the HOI classes related to the interaction class "*grab*" and "*next_to*" shown in Fig. 5 of the main paper.

5. Specific HOI splits in open set settings

In open set settings, we split the training set and testing set based on actions (predicates). Tab. 1 shows the details of the total 50 actions in VidHOI dataset. Both temporal actions and non-temporal (spatial) actions are equally splitted into training set or testing set according to the unseen percentage.

6. More results of HOI detection in closed set settings

In Tab. 2, we present more results of detecting HOIs with temporal and spatial predicates on CAD-120 dataset in closed setting, which is inaccordance with Table 4 in the main paper. The results indicate that our ACoLP model achieves better performance for both temporal-related HOIs and spatial-related HOIs compared with competing methods. Moreover, ACoLP works better on temporal-related HOIs than spatial-related HOIs due to the utilization of dynamic GNN to model temporal dynamics of action among neighboring frames.

References

[1] Meng-Jiun Chiou, Chun-Yu Liao, Li-Wei Wang, Roger Zimmermann, and Jiashi Feng. St-hoi: A spatial-temporal baseline for human-object interaction detection in videos. In *Proceedings of the 2021 Workshop on Intelligent Cross-Data Analysis and Retrieval*, pages 9–17, 2021. 3

[2] Romero Morais, Vuong Le, Svetha Venkatesh, and Truyen Tran. Learning asynchronous and sparse human-object interaction in videos. In *Proceedings of the IEEE/CVF Conference*

Non-Temporal Temporal	Train	Test
20% **unseen**	next_to, in_front_of, hold, play, ride, hug, speak_to, inside, hold_hand_of, beneath, carry, use, touch, feed, bite, drive, kiss, point_to, clean, shake_hand_with, away, towards, caress, push, pull, hit, fit, pat, grab, chase, release, wave_hand_to, shout_at, get_on, throw, get_off, open, smell, knock, lick	behind, look_at, lean_on, above, feed, cut, press, wave, squeeze, kick, close
50% **unseen**	beneath, carry, use, touch, feed, bite, drive, kiss, point_to, clean, shake_hand_with, chase, release, wave_hand_to, shout_at, get_on, throw, get_off, open, smell, knock, lick	next_to, in_front_of, hold, play, ride, hug, speak_to, inside, hold_hand_of, behind, look_at, lean_on, above, feed, cut, press, wave, squeeze, kick, close, away, towards, caress, push, pull, hit, fit, pat, grab

Table 1: Details of HOI predicates split in open set setting on VidHOI dataset. Red color indicates temporal-related actions (predicates) and green color indicates non-temporal-related actions (predicates).

Figure 1: Two examples of video HOI predictions produced by our ACoLP model. Each example consists of five consecutive keyframes. Detected person and objects are highlighted with red bounding boxes and green bounding boxes, respectively. Ground truth and predictions are listed below the frames in the form of "*interaction, object*".

on Computer Vision and Pattern Recognition, pages 16041–16050, 2021. 3

[3] Tanqiu Qiao, Qianhui Men, Frederick WB Li, Yoshiki Kubotani, Shigeo Morishima, and Hubert PH Shum. Geometric features informed multi-person human-object interaction recognition in videos. *ECCV*, 2022. 3

[4] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable human-object interaction detector with natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 939–948, 2022. 3

Method	T	S
THID, *ICDAR 2021* [4]	70.2	74.4
ST-HOI, *ICDAR 2021* [1]	71.2	74.9
ASSIGN *CVPR 2021* [2]	68.8	72.1
2G-GCN, *ECCV 2022* [3]	70.7	74.3
ACoLP, *Ours*	**77.8**	**75.2**

Table 2: HOI detection performance comparison on temporal-related (T) and static spatial-related (S) HOIs on CAD-120 dataset.

interactions	HOIs
"***grab***"	(person, grab, toy), (person, grab, hamster/rat), (person, grab, handbag), (person, grab, camera), (person, grab, snake), (person, grab, bottle), (person, grab, person), (person, grab, ball sports_ball), (person, grab, bat), (person, grab, guitar), (person, grab, cup), (person, grab, fruits), (person, grab, cat)
"***next_to***"	(person, next_to, screen monitor), (person, next_to, cup), (person, next_to, bottle), (person, next_to, table) (person, next_to, chair), (person, next_to, person), (person, next_to, dish), (person, next_to, cellphone), (person, next_to, toy), (person, next_to, ski), (person, next_to, bat), (person, next_to, guitar), (person, next_to, ball/sports_ball), (person, next_to, car), (person, next_to, dog), (person, next_to, stool), (person, next_to, electric_fan), (person, next_to, camel), (person, next_to, backpack), (person, next_to, snowboard), (person, next_to, frisbee), (person, next_to, cat), (person, next_to, handbag), (person, next_to, cake), (person, next_to, baby_seat), (person, next_to, aircraft), (person, next_to, camera), (person, next_to, crocodile), (person, next_to, penguin), (person, next_to, sink), (person, next_to, faucet), (person, next_to, hamsterrat), (person, next_to, sofa), (person, next_to, horse), (person, next_to, oven), (person, next_to, microwave), (person, next_to, snake), (person, next_to, bread), (person, next_to, crab)

Table 3: The HOI classes related to "***grab***" and "***next_ to***" in Figure 5 of the main paper.