ST-HOI: A Spatial-Temporal Baseline for
Human-Object Interaction Detection in Videos

Meng-Jiun Chiou*
National University of Singapore
Singapore, Singapore
mengjiun.chiou@u.nus.edu

Chun-Yu Liao
ASUS Intelligent Cloud Services
Taipei, Taiwan
mist_liao@asus.com

Li-Wei Wang
ASUS Intelligent Cloud Services
Taipei, Taiwan
popo55668@gmail.com

Roger Zimmermann
National University of Singapore
Singapore, Singapore
rogerz@comp.nus.edu.sg

Jiashi Feng
National University of Singapore
Singapore, Singapore
elefjia@nus.edu.sg

ABSTRACT
Detecting human-object interactions (HOI) is an important step toward a comprehensive visual understanding of machines. While detecting non-temporal HOIs (e.g., sitting on a chair) from static images is feasible, it is unlikely even for humans to guess temporal-related HOIs (e.g., opening/closing a door) from a single video frame, where the neighboring frames play an essential role. However, conventional HOI methods operating on only static images have been used to predict temporal-related interactions, which is essentially guessing without temporal contexts and may lead to sub-optimal performance. In this paper, we bridge this gap by detecting video-based HOIs with explicit temporal information. We first show that a naive temporal-aware variant of a common action detection baseline does not work on video-based HOIs due to a feature-inconsistency issue. We then propose a simple yet effective architecture named Spatial-Temporal HOI Detection (ST-HOI) utilizing temporal information such as human and object trajectories, correctly-localized visual features, and spatial-temporal masking pose features. We construct a new video HOI benchmark dubbed VidHOI where our proposed approach serves as a solid baseline.

CCS CONCEPTS
- Human-centered computing: • Computing methodologies
  Activity recognition and understanding:

KEYWORDS
Human-Object Interaction, Activity Detection, Video Understanding

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/authors.

The work was done during a research internship at ASUS Intelligent Cloud Services.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/authors.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/authors.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/authors.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/authors.

ACM Reference Format:
Meng-Jiun Chiou, Chun-Yu Liao, Li-Wei Wang, Roger Zimmermann, and Jiashi Feng. 2021. 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 (ICDAR ’21), August 21–24, 2021, Taipei, Taiwan. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3463944.3469097

1 INTRODUCTION
Thanks to the rapid development of deep learning [10, 16], machines are already surpassing or approaching human level performance in language tasks [44], acoustic tasks [46], and vision tasks (e.g., image classification [15] and visual place recognition [4]). Researchers thus started to focus on how to replicate these successes to other
Figure 2: (a) Relative performance change (in percentage), on different video tasks by replacing 2D-CNN backbones with 3D ones (blue bars) [2, 7, 36], and on VideoHOI by adding trajectory feature (tangerine bar). VideoHOI (in triplet mAP) is to detect HOI in videos and was performed ourselves on our VidHOI benchmark. STAD [11] (in triplet mAP) means Spatial-Temporal Action Detection and was performed on AVA dataset [11]. STSGG [20] (PredCls mode; in Recall@20) stands for Spatial-Temporal Scene Graph Generation and was performed on Action Genome [20]. (b) An illustration of temporal-RoI pooling in 3D baselines (e.g. [7]). Temporal pooling is usually applied to the output of the penultimate layer of a 3D-CNN (shape of $d \times T \times H \times W$) which average-pools along the time axis into shape of $d \times 1 \times H \times W$, followed by RoI Pooling to obtain feature maps of shape $d \times h \times w \times 1$. This temporal-RoI pooling, however, is equivalent to pooling the instance-of-interest feature at the same location in the keyframe throughout the video segment, which is erroneous for moving humans and objects.

Spatial-Temporal Action Detection (STAD) is another task bearing a resemblance to VideoHOI by requiring to localize the human and detect the actions being performed in videos. Note that STAD does not consider the objects that a human is interacting with. STAD is usually tackled by first using a 3D-CNN [2, 36] as the backbone to encode temporal information into feature maps. This is followed by RoI pooling with object proposals to obtain actor features, which are then classified by linear layers. Essentially, this approach is similar to a common HOI baseline illustrated in Figure 1(a) and differs only in the use of 3D backbones and the absence of interacting objects. Based on conventional HOI and STAD methods, a naive yet intuitive idea arises: can we enjoy the best of both worlds, by replacing 2D backbones with 3D ones and exploiting visual features of interacting objects? This idea, however, did not work straightforwardly in our preliminary experiment, where we replaced the backbone in the 2D baseline [39] with the 3D one (e.g., SlowFast [7]) to perform VideoHOI. The relative change of performance after replacing the backbone is presented in the left most entry in Figure 2(a) with a blue bar. In VideoHOI experiment, the 3D baseline provides only a limited relative improvement (~2%), which is far from satisfactory considering the additional temporal context. In fact, this phenomenon has also been observed in two existing works under similar settings [11, 20], where both experiments in STAD and another video task Spatial-Temporal Scene Graph Generation (STSGG) present an even worse, counter-intuitive result: replacing the backbone is actually harmful (also presented as blue bars in Figure 2(a)). We probed the underlying reason by analyzing the architecture of these 3D baselines and found that, surprisingly, temporal pooling together with RoI pooling does not work reasonably. As illustrated in Figure 2(b), temporal pooling followed by RoI pooling, which is a common practice in conventional STAD methods, is equivalent to cropping features of the same region across the whole video segment without considering the way objects move. It is not unusual for moving humans and objects in neighboring frames to be absent from its location in the target keyframe. Temporal-and-RoI-pooling features at the same location could be
getting erroneous features such as other humans/objects or meaningless background. Dealing with this inconsistency, we propose to recover the missing spatial-temporal information in VideoHOI by considering human and object trajectories. The performance change of this temporal-augmented 3D baseline on VideoHOI is represented by the tangerine bar in Figure 2(a), where it achieves ~23% improvement, in sharp contrast to ~2% of the original 3D baseline. This experiment reveals the importance of incorporating the "correctly-localized" temporal information.

Keeping the aforementioned ideas in mind, in this paper we propose Spatial-Temporal baseline for Human-Object Interaction detection in videos, or ST-HOI, which makes accurate HOI prediction with instance-wise spatial-temporal features based on trajectories. As illustrated in Figure 1(b), three kinds of such features are exploited in ST-HOI: (a) trajectory features (moving bounding boxes; shown as the red arrow), (b) correctly-localized visual features (shown as the yellow arrow), and (c) spatial-temporal actor poses (shown as the green arrow).

The contribution of our work is three-fold. First, we are among the first to identify the feature inconsistency issue existing in the naive 3D models which we address with simple yet "correct" spatial-temporal feature pooling. Second, we propose a spatial-temporal model which utilizes correctly-localized visual features, per-frame box coordinates and a novel, temporal-aware masking pose module to effectively detect video-based HOIs. Third, we establish the keyframe-based VidHOI benchmark to motivate research in detecting spatial-temporal aware interactions and hopefully inspire VideoHOI approaches utilizing the multi-modality data, i.e., video frames, texts (semantic object/relation labels) and audios.

2 RELATED WORK

2.1 Human-Object Interaction (HOI)

HOI Detection aims to reason over interactions between humans (actors) and target objects. HOI is closely related to visual relationship detection [5, 28] and scene graph generation [45], in which the subject in (subject-predicate-object) are not restricted to a human. HOI in static images has been intensively studied recently [3, 8, 13, 23–25, 29, 33, 37, 39–41, 48]. Most of the existing methods can be divided into two categories by the order of human-object pair proposal and interaction classification. The first group [3, 8, 13, 23, 24, 37, 39, 40] performs human-object pair generation followed by interaction classification, while the second group [9, 25, 29, 41] first predicts the most probable interactions performed by a person followed by associating them with the most-likely objects. Our ST-HOI belongs to the first group as we establish a temporal model based on trajectories (continuous object proposals).

In contrast to the popularity of image-based HOI, there are only a few of studies in VideoHOI [18, 22, 29, 35] and, to the best of our knowledge, all of which conducted experiments on CAD-120 [21] dataset. In CAD-120, the interactions are defined by merely temporal model based on trajectories (continuous object proposals).

In the keyframe-centric strategy. Denote $V$ as a video which has $T$ keyframes with sampling frequency of $1$ Hz as $(I_t)_t = \{1, ..., T\}$, and denote $C$ as the number of pre-defined interaction classes. Given $N$ instance trajectories including $M$ human trajectories ($M \leq N$) in a video segment centered at the target frame, for human $m \in \{1, ..., M\}$ and object $n \in \{1, ..., N\}$ in keyframe $I_t$, we aim to detect pairwise human-object interactions $r_{t,m,n} \in \{0, 1\}^C$, where each entry $r_{t,m,n,c} \in \{0, 1\}$ means whether the interaction $c$ exists or not.

2.2 Spatial-Temporal Action Detection (STAD)

STAD aims to localize actors and detect the associated actions (without considering interacting objects). One of the most popular benchmark for STAD is AVA [11], where the annotation is done at a sampling frequency of $1$ Hz and the performance is measured by framewise mean AP. We followed this annotation and evaluation style when constructing VidHOI, where we converted the original labels into the same format.

As explained in section 1, a standard approach to STAD [2, 36] is extracting spatial-temporal feature maps with a 3D-CNN followed by RoI pooling to crop human features, which are then classified by linear layers. As shown in Figure 2(a), a naive modification that incorporates RoI-pooled human/object features does not work for VideoHOI. In contrast, our ST-HOI tackles VideoHOI by incorporating multiple temporal features including trajectories, correctly-localized visual features and spatial-temporal masking pose features.

2.3 Spatial-Temporal Scene Graph Generation

Spatial-Temporal Scene Graph Generation (STSGG) [20] aims to generate symbolic graphs representing pairwise visual relationships in video frames. A new benchmark, Action Genome, is also proposed in [20] to facilitate researches in STSGG. Ji et al. [20] dealt with STSGG by combining off-the-shelf scene graph generation models with long-term feature bank [43] on top of a 2D- or 3D-CNN, where they found that the 3D-CNN actually undermines the performance. While observing similar results in VidHOI (Figure 2(a)), we go one step further to find out the underlying reason is that RoI features across frames were erroneously pooled. We correct this by utilizing object trajectories and applying Tube-of-Interest (ToI) pooling on generated trajectories to obtain correctly-localized position information and feature maps throughout video segments.

3 METHODOLOGY

3.1 Overview

We follow STAD approaches [2, 7, 36] to detect VideoHOI in a keyframe-centric strategy. Denote $V$ as a video which has $T$ keyframes with sampling frequency of $1$ Hz as $(I_t)_t = \{1, ..., T\}$, and denote $C$ as the number of pre-defined interaction classes. Given $N$ instance trajectories including $M$ human trajectories ($M \leq N$) in a video segment centered at the target frame, for human $m \in \{1, ..., M\}$ and object $n \in \{1, ..., N\}$ in keyframe $I_t$, we aim to detect pairwise human-object interactions $r_{t,m,n} \in \{0, 1\}^C$, where each entry $r_{t,m,n,c} \in \{0, 1\}$ means whether the interaction $c$ exists or not.
Correctly-localized Visual Features

We have discussed in previous sections on inappropriately pooled $N$ (including $M$ human), for each frame we utilize a trained human pose prediction model (e.g., [6]) to generate 2D actor pose feature and extract a dual spatial mask for all $M \times (N - 1)$ valid pair. The pose feature and the mask are concatenated and down-sampled, followed by two 3D convolution layers and spatial-temporal pooling to generate the masking pose features.

3.2 Correctly-localized Visual Features

Refer to Figure 1(b) for an illustration of our ST-HOI. Our model takes in a video segment (sequence of $T$ frames) centered at $h$ and utilizes a 3D-CNN as the backbone to extract spatial-temporal feature maps of the whole segment. To rectify the mismatch caused by temporal-RoI pooling, based on $N$ object (including human) trajectories $\{j_i\}, i = \{1,...,N\}, j_i \in \mathbb{R}^{T \times d}$ we generate temporal-aware features including correctly-localized features and spatial-temporal masking pose features. These features together with trajectories are concatenated and classified by linear layers. Note that we aim at a simple but effective temporal-aware baseline to VideoHOI so that we do not utilize tricks in STAD such as non-local block [42] or long-term feature bank [43], and in image-based HOI like interactiveness [24], though we note that these may be used to boost the performance.

3.3 Spatial-Temporal Masking Pose Features

Human poses have been widely utilized in image-based HOI methods [13, 24, 39] to exploit characteristic actor pose to infer some special actions. In addition, some existing works [39, 40] found that spatial information can be used to identify interactions. For instance, for human-ride-horse one can imagine the actor’s skeleton as legs widely open (on horse sides), and the bounding box center of human is usually on top of that of horse. However, none of the existing works consider this mechanism in temporal domain: when riding a horse the human should be moving with horse as a whole. We argue that this temporality is an important property and should be utilized as well.

The spatial-temporal masking pose module is presented at Figure 3(b). Given $M$ human trajectories, we first generate $M$ spatial-temporal pose features with a trained human pose prediction model. On frame $t$, the predicted human pose $h_{ijt} \in \mathbb{R}^{17 \times 2}, i = \{1,...,M\}, t = \{1,...,T\}$ is defined as 17 joint points mapped to the original image. We transform $h_{ijt}$ into a skeleton on a binary mask with $f_k : \{h_{ijt}\} \in \mathbb{R}^{17 \times 2} \rightarrow \{h_{ijt}\} \in \mathbb{R}^{1 \times H \times W}$, by connecting the joints using lines, where each line has a distinct value $x \in [0,1]$. This helps the model to recognize and differentiate different poses.

For each of $M \times (N - 1)$ valid human-object pairs on frame $t$, we also generate two spatial masks $s_{ijt} \in \mathbb{R}^{2 \times H \times W}, i = \{1,...,M\}, j = \{1,...,N - 1\}$ corresponding to human and object respectively, where
the values inside of each bounding box are ones and outsiders are zeroed-out. These masks enable our model to predict HOI with reference to important spatial information.

For each pair, we concatenate the skeleton mask \(\tilde{h}_{t, i} \) and spatial masks \(s_{t, i} \) along the first dimension to get the initial spatial masking pose feature \(p_{t, i} \in \mathbb{R}^{3 \times H \times W} \):

\[
p_{t, i} = [s_{t, i}; \tilde{h}_{t, i}].
\]

(2)

We then down-sample \(\{p_{t, i}\} \), feed into two 3D convolutional layers with spatial and temporal pooling, and flatten to obtain the final spatial-temporal masking pose feature \(\bar{p}_{t, i} \).

### 3.4 Prediction

We fuse the aforementioned features, including correctly-localized visual features \(\bar{v} \), spatial-temporal masking pose features \(\bar{p} \), and instance trajectories \(\bar{j} \) by concatenating them along the last axis

\[
\bar{v}_{so} = [\bar{v}_s; \bar{v}_o; \bar{v}_a; \bar{v}_u; \bar{v}_j; \bar{v}_{so}].
\]

(3)

where we slightly abuse the notation to denote the subscriptions \(s\) as the subject, \(o\) as the object and \(u\) as their union region. \(\bar{v}_{so}\) is then fed into two linear layers with the final output size being the number of interaction classes in the dataset. Since VideoHOI is essentially a multi-label learning task, we train the model with per-class binary cross entropy loss.

During inference, we follow the heuristics in image-based HOI [3] to sort all the possible pairs by their softmax scores and evaluate on only top 100 predictions.

### 4 EXPERIMENTS

#### 4.1 Dataset and Performance Metric

While we have discussed in section 2.1 about the problem of lacking a suitable VideoHOI dataset by analyzing CAD-120 [21], we further explain why Action Genome [20] is also not a feasible choice here. First, the authors acknowledged that the dataset is still incomplete and contains incorrect labels [19]. Second, Action Genome is produced by annotating Charades [32], which is originally designed for activity classification where each clip contains only one “actor” performing predefined tasks; should any other people show up, there are neither any bounding box nor interaction label about them. Finally, the videos are purposely-generated by volunteers, which are rather unnatural. In contrast, VidHOI are based on VidOR [30] which is densely annotated with all humans and predefined objects showing up in each frame. VidOR is also more challenging as the videos are non-volunteering user-generated and thus jittery at times. A comparison of VidHOI and the existing STAD and HOI datasets is presented in Table 1.

VidOR is originally collected for video visual relationship detection where the evaluation is trajectory-based. The volumetric Interaction Over Union (vIOU) between a trajectory and a ground truth needs to be over 0.5 before considering its relationship prediction; however, how to obtain accurate trajectories with correct start and end-timestamp remains challenging [31, 34]. We notice that some image-based HOI datasets (e.g., HICO-DET [3] and V-COCO [12]) as well as STAD datasets (e.g., AVA [11]) are using a keyframe-centered evaluation strategy, which bypasses the aforementioned issue. We thus adopt the same and follow AVA to sample keyframes at a 1 FPS frequency, where the annotations on the keyframe at timestamp \(t\) are assumed to be fixed for \(t \pm 0.5\)s. In detail, we first filter out those keyframes without presenting at least one valid human-object pair, followed by transforming the labels from video clip-based to keyframe-based to align with common HOI metrics (i.e., frame mAP). We follow the original VidOR split in [30] to divide VidHOI into a training set comprising 193,911 keyframes in 6,366 videos and a validation set with 22,808 keyframes in 756 videos. As shown in Figure 4, there are 50 relation classes including actions (e.g., push, pull, lift, etc.) and spatial relations (e.g., next to, behind, etc.). While half (25) of the predicate classes are temporal-related, they account for merely ~5% of the dataset.

Following the evaluation metric in HICO-DET, we adopt mean Average Precision (mAP), where a true positive HOI needs to meet three below criteria: (a) both the predicted human and object bounding boxes have to overlap with the ground truth boxes with IOU over 0.5, (b) the predicted target category need to be matched and (c) the predicted interaction is correct. Over 50 predicates, we follow HICO-DET to define HOI categories as 557 triplets on which we compute mean AP. By defining HOI categories with triplets we can bypass the polysemy problem [48], i.e., the same predicate word can represent very different meaning when pairing with distinct objects, e.g., person-fly-kite and person-fly-airplane. We report the mean AP over three categories: (a) Full: all 557 categories are evaluated, (b) Rare: 315 categories with less than 25 instances in the dataset, and (c) Non-rare: 242 categories with more than or equal to 25 instances in the dataset. We also examine the models in two evaluation modes: Oracle models are trained and tested with ground truth trajectories, while models in Detection mode are tested with predicted trajectories.

#### 4.2 Implementation Details

We adopt Resnet-50 [16] as our 2D backbone for the preliminary experiments, and utilize Resnet-50-based SlowFast [7] as our 3D backbone for all the other experiments. SlowFast contains the Slow and Fast pathways, which correspond to the texture details and the temporal information, respectively, by sampling video frames in different frequencies. For a 64-frame segment centered at the keyframe, \(T = 32\) frames are alternately sampled to feed into the Slow pathway; only \(T/\alpha\) frames are fed into the Fast pathway, where \(\alpha = 8\) in our experiments. We use FastPose [6] to predict human poses and adopt the predicted trajectories generated by a cascaded model of video object detection, temporal NMS and
Table 1: A comparison of our benchmark VidHOI with existing STAD (AVA [11]), image-based (HICO-DET [3] and V-COCO [12]) and video-based (CAD-120 [21] and Action Genome [20]) HOI datasets. VidHOI is the only dataset that provides temporal information from video clips and complete multi-person and interacting-object annotations. VidHOI also provides the most annotated keyframes and defines the most HOI categories in the existing video datasets.

| Dataset       | Video dataset? | Localized object? | Video hours | # Videos | # Annotated images/frames | # Objects categories | # Predicate categories | # HOI categories | Instances |
|---------------|----------------|-------------------|-------------|----------|--------------------------|---------------------|----------------------|-------------------|-----------|
| HICO-DET [3]  | ✗              | ✓                 | -           | -        | 47K                      | 80                  | 117                  | 600               | 150K      |
| V-COCO [12]   | ✗              | ✓                 | -           | -        | 10K                      | 80                  | 25                   | 259               | 16K       |
| AVA [11]      | ✓              | ✗                 | 108         | 437      | 3.7M                     | -                   | 49                   | 80                | 1.6M      |
| CAD-120 [21]  | ✓              | ✓                 | 0.57        | 0.5K     | 234K                     | 13                  | 6                    | 10                | 32K       |
| Action Genome [20] | ✓     | △                | 82          | 10K      | 157                      | 32                  | 25                   | 157               | 1.7M      |
| VidHOI        | ✓              | ✓                 | 70          | 7122     | 7.3M                     | 78†                 | 50                   | 557               | 755K      |

Figure 5: Performance comparison in predicate-wise AP (pAP). The performance boost after adding trajectory features is observed for most of the predicates. Interestingly, both spatial (e.g., next to, behind) and temporal (e.g., towards, away) predicates benefit from the temporal-aware features. Predicates sorted by the number of occurrence. Models in Oracle mode.

4.3 Quantitative Results

Since we aim to deal with a) the lack of temporal-aware features in 2D HOI methods, b) the feature inconsistency issue in common 3D HOI methods and c) the lack of a VideoHOI benchmark, we mainly compare with the 2D model [39] and its naive 3D variant on VidHOI to understand if our ST-HOI addresses these issues effectively.

Table 2: Results of the baselines and our ST-HOI on VidHOI validation set (numbers in mAP). There are two evaluation modes: Detection and Oracle, which differ only in the use of predicted or ground truth trajectories during inference. T: Trajectory features. V: Correctly-localized visual features. P: Spatial-temporal masking pose features. “%” means the full mAP change compared to the 2D model.

| Model         | Full | Non-rare | Rare | %   |
|---------------|------|----------|------|-----|
| 2D model [39] | 14.1 | 22.9     | 11.3 | -   |
| 3D model      | 14.4 | 23.0     | 12.6 | 2.1 |
| Ours-T        | 17.3 | 26.9     | 16.8 | 22.7|
| Ours-T+V      | 17.3 | 26.9     | 16.3 | 22.7|
| Ours-T+P      | 17.4 | 27.1     | 16.4 | 23.4|
| Ours-T+V+P    | 17.6 | 27.2     | 17.3 | 24.8|

Detection

| Model         | Full | Non-rare | Rare | %   |
|---------------|------|----------|------|-----|
| 2D model [39] | 2.6  | 4.7      | 1.7  | -   |
| 3D model      | 2.6  | 4.9      | 1.9  | 0.0 |
| Ours-T        | 3.0  | 5.5      | 2.0  | 15.4|
| Ours-T+V      | 3.1  | 5.8      | 2.0  | 19.2|
| Ours-T+P      | 3.2  | 6.1      | 2.0  | 23.1|
| Ours-T+V+P    | 3.1  | 5.9      | 2.1  | 19.2|

The performance comparison between our full ST-HOI model (Ours-T+V+P) and baselines (2D model, 3D model) are presented in Table 2, in which we also present ablation studies on our different features (modules) including trajectory features (T), correctly-localized visual features (V) and spatial-temporal masking pose features (P).
features (P), Table 2 shows that 3D model only has a marginal improvement compared to 2D model (overall ~2%) under all settings in both evaluation modes. In contrast, adding trajectory features (Ours-T) leads to a much larger 23% improvement in Oracle mode or 15% in Detection mode, showing the importance of correct spatial-temporal information. We also find that by adding additional temporal-aware features (i.e., V and P) increasingly higher mAPs are attained, and our full model (Ours-T+V+P) reports the best mAP in Oracle mode, achieving the highest ~25% relative improvement. We notice that the performance of Ours-T-V is close to that of Ours-T under Oracle setting, which is possibly because the ground truth trajectories (T) have provided enough "correctly-localized" information so that the correct features do not help much. We also note that the performance of Ours-T+P is slightly higher than that of Ours-T+V+P under Detection mode, which is assumably due to the same, aforementioned reason and the inferior performance resulting from the predicted trajectories. The overall performance gap between Detection and Oracle models is significant, indicating the room for improvement in trajectory generation. Another interesting observation is that Full mAPs are very close to Rare mAPs, especially under Oracle mode, showing that the long-tail effect over HOIs is strong (but common and natural).

To understand the effect of temporal features on individual predicates, we compare with predicate-wise AP (pAP) shown in Figure 5. We observe that, again, under most of circumstances naively replacing 2D backbones with 3D ones does not help video HOI detection. Both temporal predicates (e.g., towards, away, pull) and spatial (e.g., next_to, behind, beneath) predicates benefit from the additional temporal-aware features in ST-HOI. These findings verify our main idea about the essential use of trajectories and trajectory-based features. In addition, each additional features do not seem to contribute equally for different predicates. For instance, we see that while Ours-T+V+P performs the best on some predicates (e.g., behind and beneath), our sub-models achieve the highest mAP on other predicates (e.g., watch and ride). This is assumedly because predicate-wise performance is heavily subject to the number of examples, where major predicates have 10-10000 times more examples than minor ones (as shown in Figure 4).

Since the majority of HOI examples are spatial-related (~95%, as shown in Figure 4), the results above might not be suitable for demonstrating the temporal modeling ability of our proposed model. We thus focus on the performance on only temporal-related predicates in Figure 6, which demonstrates that Ours-T+V+P greatly outperforms the baselines regarding the top frequent temporal predicates. Table 3 presents the triplet mAPs of spatial- or temporal-only predicates, showing Ours-T significantly improves the 2D model on temporal-only mAP by relative +73.9%, in sharp contrast to -7.1% of the 3D model in Oracle mode. Similar to our observation with Table 2, Ours-T performs on par with Ours-T+V+P for temporal-only predicates; however, it falls short of spatial-only predicates, showing that spatial/pose information is still essential for detecting spatial predicates. Overall, these results demonstrate the outstanding spatial-temporal modeling ability of our approach.

We also compare the performance with respect to some HOI triplets in Figure 7. Similar to the results on predicate-wise mAP, we also observe the large gap between naive 2D/3D models and our models with the temporal features. ST-HOI variants are more accurate in predicting especially temporal-aware HOIs (hug/lean_on-person and push/pull-baby_walker). We also see in some examples that Ours-T+V+P does not perform the best among all the variants, e.g., lean_on-person, which is similar to the phenomenon we observed in Figure 5.

### 4.4 Qualitative Results

To understand the effectiveness of our proposed method, we visualize two video HOI examples of VidHOI predicted by the 2D model [39] and Ours-T+V+P (both in Oracle mode) in Figure 8. Each
In this paper, we addressed the inability of conventional HOI approaches to recognize temporal-aware interactions by re-focusing on neighboring video frames. We discussed the lack of a suitable setting and dataset for studying video-based HOI detection. We also identified a feature-inconsistency problem in a common video action detection baseline which arises from its improper order of RoI feature pooling and temporal pooling. To deal with the first issue, we established a new video HOI benchmark dubbed VidHOI and introduced a keyframe-centered detection strategy. We then proposed a spatial-temporal baseline ST-HOI which exploits trajectory-based temporal features including correctly-localized visual features, spatial-temporal masking pose features and trajectory features, solving the second problem. With quantitative and qualitative experiments on VidHOI, we showed that our model provides a huge performance boost compared to both the 2D and 3D baselines and is effective in differentiating temporal-related interactions. We expect that the proposed baseline and the dataset would serve as a solid starting point for the relatively underexplored VideoHOI task. Based on our baseline, we also hope to motivate further VideoHOI works to design advanced models with the multi-modal data including video frames, semantic object/relation labels and audios.

ACKNOWLEDGMENT
This research is supported by Singapore Ministry of Education Academic Research Fund Tier 1 under MOE’s official grant number T1 251RES2029.
