Interacted Object Grounding in Spatio-Temporal Human-Object Interactions

Xiaoyang Liu[1*], Boran Wen[1*], Xinpeng Liu[1,2*], Zizheng Zhou[1], Hongwei Fan[3], Cewu Lu[1], Lizhuang Ma[1], Yulong Chen[1†], Yong-Lu Li[1†]

[1] Shanghai Jiao Tong University
[2] Shanghai Innovation Institute
[3] Peking University
{liuxiaoyang0309, xinpengliu0907}@gmail.com, hwnorm@outlook.com,
{wenboran, zhou_zz, lucewu, lzma, llong_c, yonglu_li}@sjtu.edu.cn

Abstract

Spatio-temporal Human-Object Interaction (ST-HOI) understanding aims at detecting HOIs from videos, which is crucial for activity understanding. However, existing whole-body-object interaction video benchmarks overlook the truth that open-world objects are diverse, that is, they usually provide limited and predefined object classes. Therefore, we introduce a new open-world benchmark: **G**rounding **I**nteracted **O**bjects (**GIO**) including **1,098** interacted objects class and **290**K interacted object boxes annotation. Accordingly, an object grounding task is proposed expecting vision systems to discover interacted objects. Even though today's detectors and grounding methods have succeeded greatly, they perform unsatisfactorily in localizing diverse and rare objects in GIO. This profoundly reveals the limitations of current vision systems and poses a great challenge. Thus, we explore leveraging spatio-temporal cues to address object grounding and propose a 4D question-answering framework (4D-QA) to discover interacted objects from diverse videos. Our method demonstrates significant superiority in extensive experiments compared to current baselines.

Code — https://github.com/DirtyHarryLYL/HAKE-AVA

1 Introduction

As the prototypical unit of human activities, human-object interaction (HOI) plays an important role in activity understanding. Researchers begin with image-based HOI learning (Chao et al. 2018; Li et al. 2019b, 2020b; Liu et al. 2022; Wu et al. 2022) and achieve great progress. Since daily HOIs require temporal cues to avoid ambiguity in detection, e.g., *pick up-cup* and *put down-cup*, video HOI task (Damen et al. 2018; Weinzaepfel, Martin, and Schmid 2016; Zhuo et al. 2019; Materzynska et al. 2020) is proposed to advance spatiotemporal HOI (ST-HOI) learning.

However, many video HOI datasets are designed with limited predefined object classes. Charades (Sigurdsson et al. 2016), DALY (Weinzaepfel, Martin, and Schmid 2016), Action Genome (Ji et al. 2020) all have less than 50 object classes (Tab. 1). The limited object classes are less general for HOI tasks. Though some *hand-object* interactions and *egocentric* video-based HOI datasets include diverse objects like EPIC-Kitchens (Damen et al. 2018), Something-Else (Materzynska et al. 2020) and 100DOH (Shan et al. 2020), they focus on *hand-object* interactions and *egocentric* videos. As **whole body-object** interaction detection from **third-view** videos matters to numerous applications (e.g., health-care, security), here, we study third-person body-object interactions, such as *ride/sit on* (chair, horse, etc), *enter/exit* (train, bus, etc). Toward open-world HOI, we propose a large-scale third-view ST-HOI benchmark in this work, building upon AVA (Gu et al. 2018): **G**rounding **I**nteracted **O**bject (**GIO**). It contains 1,098 interacted object classes within 51 interactions and 290K frame-level triplets $\langle human, verb, object\rangle$ as Fig. 1 shows.

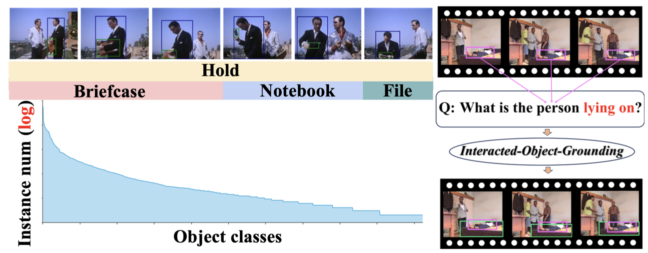


Figure 1: In daily HOIs, we interact with diverse objects with limited actions. To this end, we build GIO on AVA (Gu et al. 2018), annotating 1,000+ object classes to advance the study of HOI, with a long-tailed open-world object distribution. We propose an open-world interacted object grounding task based on GIO as in the right figure. Purple boxes indicate persons and green boxes indicate the grounded object.

Unlike previous works focussing on human/object tracking and action detection, we probed the complex ST-HOI through the **object view** given the largest scale of interacted object classes as in Fig. 1. We propose an open-world *interacted object grounding* task with corresponding metrics to formulate this challenging problem. The initial formulation of ST-HOI(Sec. 5.4) suffers from **severe missing annotation**, which makes detection and evaluation less reliable. Instead, our grounding task is insensitive to missing

---
[*]These authors contributed equally.
[†]Corresponding authors.

annotations, thus controlling the difficulty and reliability of the task, enabling a meaningful analysis. Given this task, cutting-edge image/video detectors (Ren et al. 2015; Chen et al. 2020a) fine-tuned on our train set all achieve less than 20 AP, even recent general visual grounding models based on large-scale VLMs (Liu et al. 2023) show limited performance. Hence, GIO is still challenging and essential as the touchstone for open-world HOI.

Instead of directly regressing the object box, we devise a **4D question-answering (4D-QA)** paradigm. First, the progress of the open-world segmentation model (Kirillov et al. 2023b) makes generating thorough and accurate fine-grained masks for arbitrary images possible. Then, a multi-option question-answering model is built to solve the problem: which masks correspond to the interacted object? To achieve this, multi-modal information is utilized. Besides the raw video clip, we also reconstruct the 4D human-object layout for spatial clues and take it as a representation. Despite the pixel-level accuracy of the reconstruction is limited, it is sufficient for us to tackle the occlusion and spatial ambiguities for object localization. In comparison to directly regressing the object box, the 4D human-object layout before the QA paradigm provides general object-orient HOI information, this is why our method can achieve significant improvement. We believe GIO would inspire a new line of studies and pose new challenges and opportunities for the development of deeper activity understanding.

Our contributions are three-fold: (1) We probe ST-HOI learning via an interacted object view and build a large-scale third-view ST-HOI benchmark GIO, including 290K open-world interacted object boxes from 1,098 object classes. (2) A novel interacted object grounding task is proposed to drive the studies on finer-grained activity parsing and understanding. (3) Accordingly, a 4D question-answering framework is proposed and achieves decent grounding performance on GIO with multi-modal information.

2 Related Works

Object Tracking. Object tracking is an active field and has two main branches, *i.e.*, Single-Object Tracking (Chen et al. 2020b; Fan et al. 2019) and Multi-Object Tracking (Ristani et al. 2016; Brasó and Leal-Taixé 2020). Recently, tracking-by-detection (Kim et al. 2015; Sadeghian, Alahi, and Savarese 2017) has received lots of attention and has achieved state-of-the-art performance.

Human-Object Interaction (HOI). In terms of image-based HOI learning, both image-level (Chao et al. 2015; Li et al. 2020c; Kato, Li, and Gupta 2018) and instance-level (Chao et al. 2018; Li et al. 2019b,a, 2022b, 2020a; Liu, Li, and Lu 2022) methods achieve successes with the help of large-scale datasets (Chao et al. 2018; Li et al. 2020c). As for HOI learning from third-view videos, recently many large-scale datasets (Gu et al. 2018; Sigurdsson et al. 2016; Ji et al. 2020; Shan et al. 2020; Fouhey et al. 2018; Caba Heilbron et al. 2015) are released to promote this field, thus providing a data basis for us. They provide clip-level (Caba Heilbron et al. 2015; Fouhey et al. 2018; Sigurdsson et al. 2016) or instance-level (Gu et al. 2018; Ji et al. 2020; Weinzaepfel, Martin, and Schmid 2016) action labels, but few of them afford diverse object classes. Though some datasets (Materzynska et al. 2020; Damen et al. 2018) provide instance labels of diverse object classes, they usually concentrate on egocentric hand-object interaction understanding (Xu, Li, and Lu 2022). Relatively, we focus on whole-body-object interaction learning based on third-view videos and propose GIO featuring the discovery of diverse objects. Recently, there are also methods studying video-based visual relationship (Shang et al. 2017; Liu et al. 2020) and HOI (Qi et al. 2018; Wang and Gupta 2018; Baradel et al. 2018; Girdhar et al. 2019).

Object Detection and Localization. Object detection (Ren et al. 2015; Redmon et al. 2016) achieves huge success with deep learning and large-scale datasets (Lin et al. 2014) but may struggle without enough training data. Some works (Fan et al. 2020) study few/zero-shot detection. Moreover, as videos can provide temporal cues of moving objects, video object detection (Chen et al. 2020a) also received attention. Unlike typical detection, some studies try to utilize context cues, such as human actor (Kim et al. 2020; Gkioxari et al. 2018), action recognition (Yuan et al. 2017; Yang et al. 2019), object relation (Hu et al. 2018), to advance object localization. Gkioxari et al. (2018) treated object localization as density estimation and used a Gaussian function to predict object location. Kim et al. (2020) borrowed human pose cues and language prior, constructing a weakly-supervised detector. Moreover, object grounding with language descriptions also attracts attention in the vision-language crossing field, with promising potential in open-vocabulary object detection. Li et al. (2022a) formulates object detection as an object grounding problem for open-vocabulary object detection. Yao et al. (2022) boosted data from image captioning datasets for generalization ability. Liu et al. (2023) extended the powerful DINO (Zhang et al. 2022) model for the object grounding pipeline, achieving impressive performance. Sadhu, Chen, and Nevatia (2020) grounded objects in video clips given language descriptions.

3 Constructing GIO

3.1 Data Collection

To support practical ST-HOI learning, we collect third-view videos from large-scale dataset AVA (Gu et al. 2018). It contains 430 videos with spatio-temporal labels of 80 atomic actions (body motions and HOIs). As AVA includes complex HOIs in diverse scenes, it can bring great visual diversity to our benchmark. We extract the HOI-related frames and the corresponding human boxes and action labels, thus the clips in GIO have uneven temporal durations. Notably, we only consider the non-human objectives in HOIs. Overall, based on the available train and validation (val) sets of AVA 2.2 (Gu et al. 2018) (299 videos), we chose 74 hours of video including 51 actions (detailed in the supplementary).

3.2 Dataset Annotation

AVA provides labels with a stride of 1s, so we add boxes and class labels for all interacted objects with the same stride. Following AVA, we define the annotated frame as key frames which are at 1-second intervals.

Dataset	Video Hours	Annotated Frames	Objects		HOI		HOI/frame	View	Subjective
			class	instance	class	triplet			
Something-Something (Goyal et al. 2017)	121	108K	-	-	174	-	-	first	hand
100DOH (Shan et al. 2020)	3144	100K	-	110.1K	5	189.6K	1.90	first, third	hand
Something-Else (Materzynska et al. 2020)	-	8M	18K*	10M	174	6M	0.75	first	hand
EPIC-Kitchens (Damen et al. 2018)	55	266K	331	454K	125	243K	0.91	first	hand
CAD120++ (Zhuo et al. 2019)	0.57	61K	13	64K	10	32K	0.52	third	head, hand
VLOG (Fouhey et al. 2018)	344	114K	30	-	9	-	-	first, third	hand
AVA (Gu et al. 2018)	107	351K	-	-	51	-	-	third	whole body
Charades (Sigurdsson et al. 2016)	82	66K	46	41K	30	-	-	third	whole body
DALY (Weinzaepfel, Martin, and Schmid 2016)	31	11.9K	43	11K	10	11K	0.92	third	whole body
Action Genome (Ji et al. 2020)*	82	234K(227K*)	35	476K	15*	454K*	2.01	third	whole body
VidHOI (Chiou et al. 2021)*	70	217K(146K*)	78	-	39*	278K*	1.90	third	whole body
GIO	74	126K	**1,098**	290K	51	290K	**2.30**	third	whole body

Table 1: Dataset comparison. Instances/triplets are in *frame-level*. *18K**: object class labels of Materzynska et al. (2020) are uncurated. In *Action Genome** and *VidHOI**, spatial relationships are not regarded as HOI.

First, as humans can perform multi-interaction simultaneously, we set the **annotating unit** as a clip including one *single* interaction to normalize the annotation. For example, a 30s clip including an actor *holds-sth* (1-30s) and *inspects-sth* (10-15s), will be divided into two sub-clips, i.e., a 30s sub-clip for *holds-sth* and a 5s sub-clip for *inspects-sth*. In brief, each sub-clip contains **one** verb and **one/several** class-agnostic interacted objects. Then, sub-clips are annotated separately, and each one is annotated by at least 3 annotators and checked by an expert to ensure quality.

Second, as AVA contains various scenarios and diverse objects, to better locate objects and avoid ambiguity, each annotator is given a whole sub-clip to draw boxes and classify them. In default, we use COCO (Lin et al. 2014) 80 objects as a class pool. If annotators think an object is not in the pool, they are asked to input a suitable class according to their judgments. If an object cannot be recognized, they can choose the "unknown" option. Then, we find that a surprising 42.66% of object instances are beyond our pool. After exhaustive annotation, we fix the input typos, exclude outliers via clustering, and combine similar items. Finally, 1,098 classes are extracted after cleaning. We then conduct re-recognition for the frames including "unknown" objects.

Finally, to generate the ST-HOI labels, we further consider the objects in each sub-clip (one interaction of one person). If there is only one object in a sub-clip, we use its locations as the labels. If there are multiple objects, we record all of their boxes and manually link their boxes as multiple-object tracklets. Then, each sub-clip is seen as a *ST-HOI traklet*, whose label records *a human actor tracklet, an interaction, a/several class-agnostic object tracklets*.

3.3 Dataset Statistics and Attributes

GIO includes 290K HOI triplets and 290K object boxes of 1,098 classes. Only **20.85%** of our object classes are covered by the large-scale object dataset FSOD (Fan et al. 2020). It is noteworthy that Action Genome and VidHOI include predicates such as `next to`, which are not HOIs. Consequently, we recalculated the statistics in Tab. 1. In contrast, GIO, aiming for diversity and finer granularity, offers the most object classes and the richest HOI instances per frame (**2.30**).

3.4 Interacted Object Grounding

GIO supports ST-HOI detection and fine-grained tasks like object classification. The original ST-HOI task, involving detection, tracking, and action recognition, is complex and challenging due to inherent difficulties and annotation quality. So instead of requiring models to detect complete ST-HOI triplets, we focus on GIO's capability for interacted object grounding, i.e., given the human actor tracklet (and the interaction semantics), while object labels are not included in the interaction semantics, probing the ST-HOI understanding from the object view. In GIO, 328 object classes only have less than 5 samples (boxes) in our train set, and 98 classes are *unseen* in the inference.

4 Method

In this section, we describe the pipeline of our method (Fig. 2). We focus on interacted object grounding, i.e., given the human actor tracklet (and the interaction semantics), systems are required to ground the interacted object. The difference between our task and the common object grounding tasks is our focus on the specific interaction between the grounded object and the person (*interactiveness*), which makes it more difficult. For clarity, the description unit hereinafter is *one human tracklet* including one tracked person.

4.1 Overview

Given a clip C, the human tracklet $T_h = \{I_h^k\}_{k=1}^n$ (n for tracklet length), we aim at learning a model $\mathcal{M}$ as $\hat{T_o} = \mathcal{M}(C, T_h, \{s, \emptyset\})$. $\hat{T_o}$ is the predicted interacted object tracklet and the semantics s is an *optional* input to inform the system with high-level semantics. To this end, we adopt a novel 4D question-answering paradigm to leverage HOI prior. Given the strong ability of SAM (Kirillov et al. 2023a), we adopt it as an objectness detector to generate candidate object proposals (Sec. 4.2). The clip C is first fed to SAM, resulting in K candidate object mask tracklets $M_o = \{M_o^i\}_{i=1}^K$. The task is then reformulated as choosing the interacted mask tracklets from the candidate tracklets, as

$$\hat{T_o} = \mathcal{M}(C, T_h, \{s, \emptyset\}, M_o). \quad (1)$$

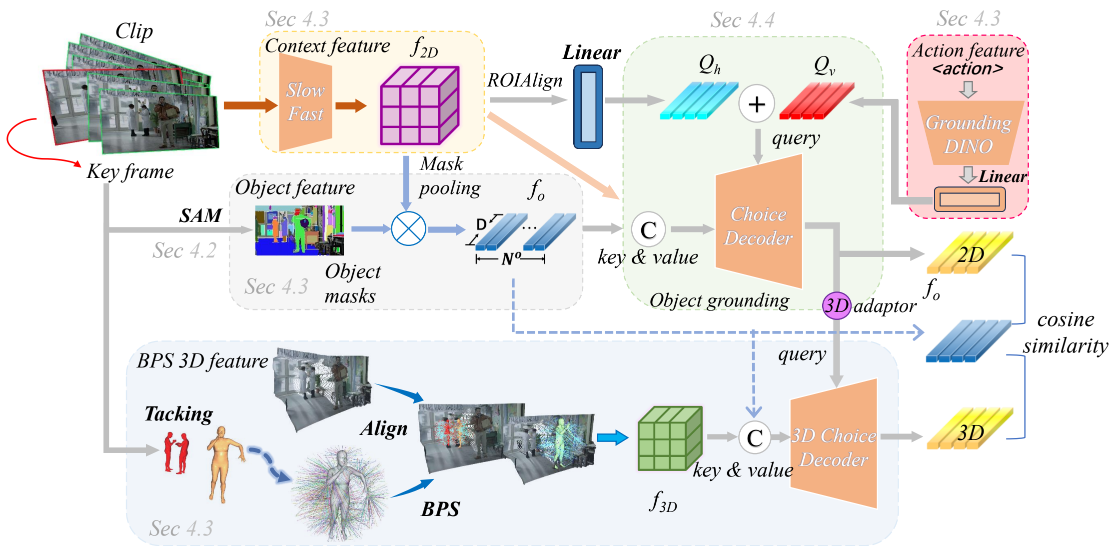


Figure 2: The overview of our 4D-QA. It utilizes a 4D question-answering paradigm to effectively locate the interacted objects.

To tackle the challenging GIO, a 4D question-answering network is devised as shown in Fig. 2. Multimodal features, including 4D clues, are extracted in the inspiration of DJ-RN (Li et al. 2020a). We begin by extracting spatiotemporal features from the video using the SlowFast (Feichtenhofer et al. 2019) network as a basis. Then, the 4D Human-Object layout is reconstructed for feature extraction (Sec. 4.3). Finally, we ground the interacted object with two decoders to summarize the important clues in complex spatiotemporal patterns (Sec. 4.4). Despite the suboptimal precision of 4D Human-Object reconstruction, it is effective in alleviating the *view ambiguity* in clips, also enhancing the object localization with 3D spatial information. The question-answering paradigm eases the learning process.

4.2 SAM-based Candidate Generation

We chose SAM as the candidate proposal generator for several reasons. First, SAM, based on pixel-level segmentation, provides a finer granularity and more accurate segmentation. Second, AVA consists of many video scenes that are dark, complex, and contain numerous objects. Traditional detection methods struggle to accurately predict small and blurry objects in such challenging scenarios. In contrast, SAM's pixel-based segmentation is more robust and accurate than directly predicting object bounding boxes. In addition, SAM is also adept at dealing with large objects. However, SAM could segment objects into multiple parts. Thus, our policy is to predict vast majority of the masks belong to the object resulting in a highly accurate bounding box.

Mask proposal generation. Given a clip C, we denote the keyframe as C_k. SAM is first fed with a grid of point prompts on C_k. Then, low-quality and duplicate masks are filtered out. As a result, each image would produce at most 255 masks as M_o, which will be sent to the model as proposals to generate the final object box.

(a) Original image. (b) SAM masks.

(c) Accurate bbox. (d) GT mask&bbox.

Figure 3: SAM-based candidate generation.

GT proposals. To judge which mask is GT, we input the GT object box to SAM as the prompt to get an accurate mask (M_{acc}). Next, we calculate the area of the intersection between the *proposal* masks and the *accurate* mask (A_{inter}), and divide them by the area of the proposal masks A_p to get a ratio for each proposal mask as $ratio = A_{inter}/A_o$. Masks with a ratio greater than 0.9 are identified as GT masks. Fig. 3 demonstrates the above process.

4.3 Multi-Modal Feature

To fully leverage the temporal and spatial continuity features of videos, including object information, HOI details, and spatial relations from multiple views, we employed a multi-modal feature extraction approach.

Context Feature. We utilize widely-used SlowFast (Feichtenhofer et al. 2019) to extract context features from the video clip C. The features from slow and fast branches are pooled along the time axis, then concatenated into the context feature map $f_c \in \mathcal{R}^{H\times W\times D}$, with $H \times W$ being the feature map resolution, and D is the feature dim.

Object Feature. We first resize the i-th mask M_o^i concerning the context feature map, then the feature could be computed as $f_o^i = AvgPool(f_c \otimes M_o^i) \in \mathcal{R}^D$, where $\otimes$ indicates element-wise multiplication. The object feature is denoted as $f_o \in \mathcal{R}^{N_o\times D}$ with N_o masks.

Language Interaction Feature is optional. If adopted, we input the language-guided query embedding $f_v \in \mathcal{R}^D$ of GroundingDINO (Liu et al. 2023), which needs a language prompt and the key frame as input. Some other interaction features are discussed in Sec. 5.6.

4D Human-Object Feature. Inspired by Li et al. (2020a), which utilizes 3D information for HOI learning, we incorporate 3D information into our pipeline to exploit the rich HOI prior carried by 4D information. Specifically, we lift GIO to 4D by reconstructing the HOIs in 3D. However, lifting GIO to 4D is challenging given its diverse objects. Existing efforts usually require 3D templates for the objects, which is inapplicable for open-world GIO. To alleviate this, we adopt depth estimation for holistic scene estimation, bypassing the need for object templates. Then, we align the human and scene for consistent 4D H-O representation. Finally, we extract the 3D feature with the lightweight base point set (BPS) (Prokudin, Lassner, and Romero 2019).

1) **Human reconstruction.** Considering that videos *without scene switching* allow for better human tracking and less processing time of 3D data, we first perform shot detection and segment the original video into multiple sub-clips. Then, PHALP (Rajasegaran et al. 2022) is adopted to recover 4D human tracklets from the sub-clips in SMPL (Loper et al. 2015) representation. The 3D humans are further represented as SMPL mesh point clouds $p_h \in \mathcal{R}^{T\times N_h\times V\times 3}$, where T is the length of the clip, N_h is the number of existing human instances, and V is the number of mesh vertices.

2) **Scene reconstruction via depth estimation**. We use ZoeDepth (Bhat et al. 2023) to estimate the depth of the corresponding clip and transform them into scene point cloud $p_s \in \mathcal{R}^{T\times N_p\times 3}$, where N_p is the number of points.

3) **Human-Scene alignment**. The humans and scenes are initially inconsistent in scale and position. To align them, we render the N_f front surface vertices $p_h^f \in \mathcal{R}^{N_f\times 3}$ of the human mesh to the image space, find the corresponding pixel of each vertice, and locate the corresponding point in the scene point cloud $p_s^f \in \mathcal{R}^{N_f\times 3}$. Next, we align p_h^f and p_s^f by calculating the scale and displacement of p_s^f to align with p_h^f. We calculate scale s and displacement b as

$$\begin{aligned}
d_h &= \frac{1}{{N_f}^2}\sum_{i=1}^{N_f}\sum_{j=1}^{N_f}\|p_{h\,i}^f - p_{h\,j}^f\|_2, \\
d_s &= \frac{1}{{N_f}^2}\sum_{i=1}^{N_f}\sum_{j=1}^{N_f}\|p_{s\,i}^f - p_{s\,j}^f\|_2, \\
s &= \frac{d_h}{d_s}, p_s^f * = s, b = \frac{1}{N_f}\sum_{i=1}^{N_f} p_{h\,i}^f - \frac{1}{N_f}\sum_{j=1}^{N_f} p_{s\,j}^f.
\end{aligned} \tag{2}$$

In detail, the scale is calculated as the ratio between the average pairwise distance of p_h^f and p_s^f, while the displacement is calculated as the displace between the center point of p_h^f and p_s^f. The aligned human-scene point cloud is then formulated as $p = (p_h, p_s \cdot s + b) \in \mathcal{R}^{T\times(N_h\times V + N_h\times N_p)\times 3}$.

4) **3D feature extraction.** We adopt BPS to extract features, which is simple and efficient for encoding 3D point clouds into fixed-length representations. We randomly select $\frac{D}{2}$ fixed points in a sphere and compute vectors from these basis points to the nearest points in a point cloud; then use these vectors (or simply their norms) as features, shown in Fig. 2. We adopt the human pelvis joint as the sphere center for base point generation. We selected a radius of 1.5 times the height of the human body to cover the range of human interactions. In this way, in one space, we obtain $T \times N_h \times \frac{D}{2}$ base points. We calculate the distances from these base points to the human mesh point cloud and the scene point cloud, treating them as features. Then we concatenate human features and scene features to get the final 3D feature $f_{3D} \in \mathcal{R}^{(T\times N_h)\times D}$, in the following we refer to $T \times N_h$ by N_{3D}, i.e., $\mathcal{R}^{N_{3D}\times D}$.

4.4 Object Grounding

We utilize a 2D transformer decoder and a 3D transformer decoder to integrate multi-modal features. The 2D decoder outcome is sent to the 3D decoder as a query via an MLP as the 3D adapter. Note that the 2D decoder results have already been satisfactory, but the 3D decoder could further enhance predictions from the 3D perspective. Each 2D decoder query $Q_s \in \mathcal{R}^{N_q\times D}$, is obtained via $Q_s = Q_v + Q_h$, where $Q_v \in \mathcal{R}^{N_q\times D}$ is the optional verb semantic query from the feature vector f_v, and the human query $Q_h \in \mathcal{R}^{N_q\times D}$ is obtained via a temporal pooling, a ROIAlign pooling, and a spatial pooling of the SlowFast features with the human bounding box. Given the context feature f_c, the object feature f_o, we concatenate them as the key and value of the 2D decoder. The object feature f_o and the 3D feature f_{3D} are concatenated as the key and value of the 3D decoder.

The 2D/3D decoder outputs feature f_q. The cosine similarity between f_q and all object mask features f_o is computed. Then, we derive scores for each query relative to each mask, denoted as S_m^i. Higher scores suggest a greater likelihood of the mask being associated with the target object. Considering that a person tends to interact with objects that are closer in proximity, we use the distance between masks and humans to assist us in calculating mask scores. The distance of the i-th mask is computed as $S_d^i = dist(C_h, C_m^i)$, where C_h and C_m^i refer to the human box and the i-th mask's box. Ultimately we adopt the GIoU (Rezatofighi et al. 2019) distance. The final score of the i-th mask is computed as

$$S_f^i = \gamma \times S_m^i + (1-\gamma) \times S_d^i, \tag{3}$$

where γ is a weight. We introduce a threshold τ to determine whether a mask belongs to the target object. If no mask scores exceed the threshold, we select the highest-scoring mask. We then cluster the predicted masks by depth and define the boundaries (see supplementary material). For the i-th mask of an object, BCE loss L_o^i is used for supervision. The overall loss is computed as $L_o = (\sum_{i=1}^{N_o} L_o^i)/N_o$.

5 Experiments

5.1 Setting

Modified versions of mean Average Precision (mAP) and mean Intersection over Union (mIoU) are adopted. For each GT tracklet, we sort all predictions by their scores in descending order. We identify the first prediction with an IoU higher than a threshold as a hit and calculate its precision by its position in that order. mAP is averaged across all test instances. For mIoU, we calculate all IoUs between the GT and predicted boxes and report the largest IoU. To take into account the precision of the prediction, a *weighted* mIoU is proposed as $\mathrm{mIoU_w}$. For each GT tracklet, predictions are sorted by scores in descending order. The rank of each prediction is used to calculate $\mathrm{mIoU_w}$ as

$$mIoU_w(T_o) = \frac{\sum_{\hat{T_o}} IoU(T_o, \hat{T_o})/rank(\hat{T_o})}{\sum_{\hat{T_o}} 1/rank(\hat{T_o})}, \quad (4)$$

where $\hat{T_o}$ and T_o denotes predicted and GT tracklets. Since $\mathrm{mIoU_w}$ is a more reasonable metric, we adopt $\mathrm{mIoU_w}$ instead of mIoU in the experiments.

5.2 Implementation Details

For the 3D feature, considering the reconstruction quality of the 4D HOI layout, the reconstruction is only conducted for frames with object labels. After filtering, there are 107,663 of 126,700 key-frames attached with 4D HOI layout (85,370 for training, 22,293 for inference). SlowFast pre-trained on AVA 2.2 is adopted for video feature extraction. An Adam optimizer, an initial learning rate of 1e-3, a cosine learning rate schedule, and a batch size of 16 are adopted. When training the 2D decoder, the learning rate of the parameters of SlowFast and Grounding DINO is 1e-5 and the 3D decoder is omitted. When training the 3D decoder, other parts except the 3D decoder are frozen. N_{3D} is set to 256, N_o is set to 256 and N_q is set to 24 for alignment. Considering that the ground truth mask for each keyframe is sparse, we use weighted BCE loss, where the loss coefficient for true positions is ten times that of false positions.

5.3 Baselines

We adopt six models of four different types as our baseline. It is worth mentioning that since our task is new, we find these models most close to our task. But, they still do not fit our task very well in the setting. We devise corresponding protocols to adapt these models to our task.

Image/Video-based HOI models. PViC (Zhang et al. 2023) and Gaze (Ni et al. 2023) are adopted as conventional image/video-based HOI detection baselines. Given a frame or clip and a human bounding box b, the HOI models input the frame or clip and output a series of HOI triplets as $\langle b_h, b_o, p \rangle$, where b_h, b_o are human and object bounding boxes, and p is the predicted interaction probability. We preserve all the results with $IoU(b, b_h) > 0.5, p > 0.2$, and the corresponding b_o are adopted as the grounded objects.

Open-vocabulary object detection models. Detic (Zhou et al. 2022) is adopted, inputting a frame and expected object categories, outputting $\langle b_o, p \rangle$ as object bounding boxes and objectness score. Results with $p > 0.5$ are preserved and paired with the human query as the grounded objects.

Methods	mAPs					mIoU_w
	@0.5	@0.6	@0.7	@0.8	@0.9	
PViC (Zhang et al. 2023)	11.78	9.89	7.73	5.31	2.45	11.64
Gaze (Ni et al. 2023)	12.71	8.18	5.43	3.14	1.17	16.06
Detic (Zhou et al. 2022)	12.17	7.63	4.89	2.85	1.10	13.91
CG-STVG (Gu et al. 2024)	12.35	8.97	6.17	3.70	1.72	17.50
Grounding DINO (Liu et al. 2023)	17.53	12.86	9.43	6.15	2.73	20.41
Qwen-VL (Bai et al. 2023)	14.12	8.83	5.39	2.91	1.11	20.28
4D-QA	**23.38**	**18.48**	**14.40**	**10.71**	**6.39**	**29.71**

Table 2: Results on GIO with multiple baselines.

Visual grounding models. Grounding DINO (Liu et al. 2023) is adopted, which takes a frame and a text prompt s as input and produces $\langle b_o, p \rangle$ as grounded box and confidence. We also test a video-grounding baseline CG-STVG (Gu et al. 2024), which aims to predict a spatial-temporal tube for a specific target subject/object given some semantic s. s is in the format as "The object that the person is {interacting with}", where the placeholder "{Interacting with}" could be replaced with a specific action name. All the outputs are paired with the human query as the grounded object.

LLM based models. Qwen-VL (Bai et al. 2023) is adopted. It takes a frame and the text prompt "Output the bounding box that the person is {interacting with}." and produces the bounding box b_o if detected.

5.4 Results

Results are shown in Tab. 2. As HOI detection models, PViC and Gaze fail to perform well due to the large number of novel objects in GIO. The open-vocabulary object detection model Detic demonstrates low $mIoU_w$ since it cannot discriminate the interacted objects related to humans (interactiveness (Li et al. 2019b)). It is noteworthy that Detic tends to predict a substantial number of object bounding boxes, sometimes exceeding 900, with many false positive predictions. CG-STVG, lacking pre-training on large datasets and integration of visual-language models, outperforms PViC, Gaze, and Detic in mIoU_w, using a single high-quality bounding box per HOI instance for higher mIoU_w despite lower mAP. Grounding DINO performs better than other baselines, but it is still limited for "hidden objects". Also, it frequently fails to fully understand the interaction semantics. Qwen-VL, a large vision language model, provides decent $mIoU_w$ but poor mAPs, which suggests that although Qwen-VL can localize the approximate positions of most objects, it struggles to detect precise bounding boxes. Our model performs well in localizing diverse and unseen objects, where the baselines struggle. Also, our model demonstrated decent $mIoU_w$. These experimental findings indicate that our method excels in object grounding for spatiotemporal HOI understanding.

In addition, we considered ST-HOI as the task design, resulting in the highest mAP of **6.8**, i.e., the ST-HOI task is kind of **too challenging** even ignoring the annotation missing problem. This demonstrates the rationality and exploratory potential of the GIO task formulation.

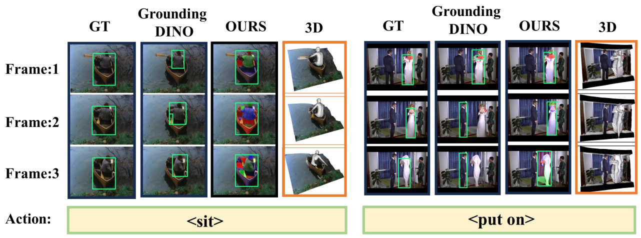
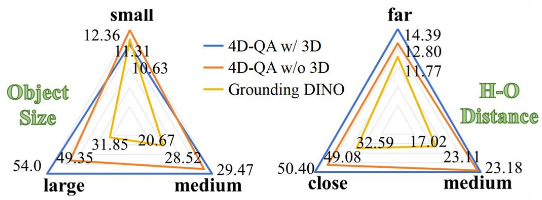
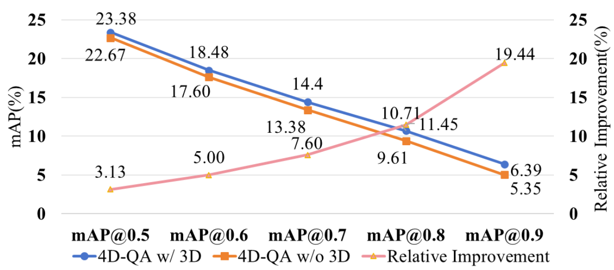


Figure 4: Visualization of interacted object grounding. We also list the reconstructions.



(a) mIoU$_w$ w.r.t. object size and H-O distance.



(b) mAP w.r.t. IoU threshold.

Figure 5: Fine-grained performance analysis.

We also evaluate 4D-QA on VidHOI (Chiou et al. 2021) with the GIO task. The GIO-pretrained 4D-QA gets 14.23 mAP and 22.66 mIoU$_w$ and **25.35** mAP and **29.61** mIoU$_w$ after 10 epoch finetuning. GroundingDINO gets **15.87** mAP and **17.57** mIoU$_w$. Results on VidHOI reveal that *grounding interacted objects* is challenging enough to be carefully explored and our 4D-QA maintains decent performance. VidHOI only has **1.22** HOIs/frame in the filtered (Sect. 3.3) test set, allowing Grounding DINO to slightly outperform zero-shot 4D-QA on mAP because Grounding DINO performs better to localize the only object in the one-HOI frame.

5.5 Visualization

Fig. 4 visualizes the grounded interacted object in 3 consecutive frames. The predicted masks (colored regions) are integrated into final object boxes (green) as in Sec. 4.4.

5.6 Ablation Study

We conduct ablation studies on the different components of our model on the GIO test set as reported in Tab. 3.

Distance. Removing the use of distance would result in a degradation (**22.07** mAP, **27.85** mIoU$_w$). Replacing the GIoU distance with the L2 distance would also cause a decline in performance (**22.34** mAP, **28.39** mIoU$_w$).

Model	mAP@0.5	mIoU$_w$
4D-QA	**23.38**	**29.71**
4D-QA w/o 3D	22.67	28.76
4D-QA w/ L_2 distance	22.34	28.39
4D-QA w/o distance	22.07	27.85
4D-QA w/ Bert	20.02	28.11
4D-QA w/ CLIP	20.00	28.15
4D-QA w/o action	19.04	27.30
4D-QA w/ pred hbox w/o action	18.92	27.17
4D-QA w/ box regression	18.82	26.88

Table 3: Ablation Results.

3D Feature f_{3D}. 4D-QA w/o f_{3D} presents a performance decrease (**22.67** mAP, **28.76** mIoU$_w$). Fig. 5(a) further shows the influence of f_{3D} w.r.t. different data characteristics, namely the relative object size and H-O distance. As shown, our methods provide superior performance compared to Grounding DINO across different data groups, especially on medium-to-large objects and close-to-medium H-O pairs. For 2D and 3D's difference, f_{3D} gains **4.68** mIoU$_w$ on large objects and **1.32** mIoU$_w$/**1.59** mIoU$_w$ on close/far H-O pairs. Besides, we find f_{3D} outperforms 2D by **2.25** mIoU$_w$ and **2.83** mAP@0.5 on **40** verb classes, especially on verbs like `drive`, `exit`, `press` that involve large or occluded objects or occur in complex scenarios. Notably, in Fig. 5(b) the relative improvement that f_{3D} brings increases with the IoU threshold requirement for mAP, indicating that f_{3D} contains more accurate predictions than 2D.

Interaction Feature. We replaced the Grounding DINO interaction feature with the Bert and CLIP interaction language embedding. In addition, we performed a test without the interaction feature. As shown, the sophisticated feature from the Grounding DINO can help the grounding task to better utilize the interaction information, while the simple language representation difference between Bert and CLIP affects the performance little. Eliminating the interaction feature brings a major performance degradation.

Predicted human boxes, with 88 mIoU w.r.t. GT human boxes, were used as human queries. The slight performance drop indicates the robustness and flexibility of 4D-QA.

Box Regression. We used an MLP after the decoder to directly regress the boxes instead of utilizing SAM mask candidates or other box proposals. The performance drop shows the importance of SAM-generated mask candidates.

6 Conclusion

We constructed GIO, which consists of many rare objects that are overlooked but important in HOI learning. 290K frame-level HOI triplets annotations with 1,098 objects were collected. Based on GIO, an interacted object grounding task was devised and a 4D-QA framework was proposed to tackle this challenging task with decent results. We believe GIO would inspire deeper activity understanding and interactive object grounding, thus enhancing the performance of tasks associated with spatiotemporal analysis and exploration.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China under Grants No.62302296, 62306175, and 62472282.

References

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966.

Baradel, F.; Neverova, N.; Wolf, C.; Mille, J.; and Mori, G. 2018. Object level visual reasoning in videos. In *ECCV*.

Bhat, S. F.; Birkl, R.; Wofk, D.; Wonka, P.; and Müller, M. 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*.

Brasó, G.; and Leal-Taixé, L. 2020. Learning a neural solver for multiple object tracking. In *CVPR*.

Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*.

Chao, Y.-W.; Liu, Y.; Liu, X.; Zeng, H.; and Deng, J. 2018. Learning to Detect Human-Object Interactions. In *WACV*.

Chao, Y. W.; Wang, Z.; He, Y.; Wang, J.; and Deng, J. 2015. HICO: A Benchmark for Recognizing Human-Object Interactions in Images. In *ICCV*.

Chen, Y.; Cao, Y.; Hu, H.; and Wang, L. 2020a. Memory Enhanced Global-Local Aggregation for Video Object Detection. In *CVPR*.

Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; and Ji, R. 2020b. Siamese Box Adaptive Network for Visual Tracking. In *CVPR*.

Chiou, M.-J.; Liao, C.-Y.; Wang, L.-W.; Zimmermann, R.; and Feng, J. 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*, 9–17.

Damen, D.; Doughty, H.; Maria Farinella, G.; Fidler, S.; Furnari, A.; Kazakos, E.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*.

Fan, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Bai, H.; Xu, Y.; Liao, C.; and Ling, H. 2019. Lasot: A high-quality benchmark for large-scale single object tracking. In *CVPR*.

Fan, Q.; Zhuo, W.; Tang, C.-K.; and Tai, Y.-W. 2020. Few-Shot Object Detection with Attention-RPN and Multi-Relation Detector. In *CVPR*.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *ICCV*.

Fouhey, D. F.; Kuo, W.-c.; Efros, A. A.; and Malik, J. 2018. From lifestyle vlogs to everyday interactions. In *CVPR*.

Girdhar, R.; Carreira, J.; Doersch, C.; and Zisserman, A. 2019. Video action transformer network. In *CVPR*.

Gkioxari, G.; Girshick, R.; Dollár, P.; and He, K. 2018. Detecting and recognizing human-object interactions. In *CVPR*.

Goyal, R.; Kahou, S. E.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. 2017. The" Something Something" Video Database for Learning and Evaluating Visual Common Sense. In *ICCV*.

Gu, C.; Sun, C.; Ross, D. A.; Vondrick, C.; Pantofaru, C.; Li, Y.; Vijayanarasimhan, S.; Toderici, G.; Ricco, S.; Sukthankar, R.; Schmid, C.; and Malik, J. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*.

Gu, X.; Fan, H.; Huang, Y.; Luo, T.; and Zhang, L. 2024. Context-Guided Spatio-Temporal Video Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18330–18339.

Hu, H.; Gu, J.; Zhang, Z.; Dai, J.; and Wei, Y. 2018. Relation networks for object detection. In *CVPR*.

Ji, J.; Krishna, R.; Fei-Fei, L.; and Niebles, J. C. 2020. Action Genome: Actions As Compositions of Spatio-Temporal Scene Graphs. In *CVPR*.

Kato, K.; Li, Y.; and Gupta, A. 2018. Compositional learning for human object interaction. In *ECCV*.

Kim, C.; Li, F.; Ciptadi, A.; and Rehg, J. M. 2015. Multiple hypothesis tracking revisited. In *ICCV*.

Kim, D.; Lee, G.; Jeong, J.; and Kwak, N. 2020. Tell Me What They're Holding: Weakly-Supervised Object Detection with Transferable Knowledge from Human-Object Interaction. In *AAAI*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023a. Segment Anything. *arXiv:2304.02643*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023b. Segment anything. *arXiv preprint arXiv:2304.02643*.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022a. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10965–10975.

Li, Y.-L.; Liu, X.; Lu, H.; Wang, S.; Liu, J.; Li, J.; and Lu, C. 2020a. Detailed 2D-3D Joint Representation for Human-Object Interaction. In *CVPR*.

Li, Y.-L.; Liu, X.; Wu, X.; Huang, X.; Xu, L.; and Lu, C. 2022b. Transferable Interactiveness Knowledge for Human-Object Interaction Detection. In *TPAMI*.

Li, Y.-L.; Liu, X.; Wu, X.; Li, Y.; and Lu, C. 2020b. HOI Analysis: Integrating and Decomposing Human-Object Interaction. In *NeurIPS*.

Li, Y.-L.; Xu, L.; Liu, X.; Huang, X.; Xu, Y.; Chen, M.; Ma, Z.; Wang, S.; Fang, H.-S.; and Lu, C. 2019a. Hake: Human activity knowledge engine. *arXiv preprint arXiv:1904.06539*.

Li, Y.-L.; Xu, L.; Liu, X.; Huang, X.; Xu, Y.; Wang, S.; Fang, H.-S.; Ma, Z.; Chen, M.; and Lu, C. 2020c. PaStaNet: Toward Human Activity Knowledge Engine. In *CVPR*.

Li, Y.-L.; Zhou, S.; Huang, X.; Xu, L.; Ma, Z.; Fang, H.-S.; Wang, Y.; and Lu, C. 2019b. Transferable interactiveness knowledge for human-object interaction detection. In *CVPR*.

Lin, T. Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*.

Liu, C.; Jin, Y.; Xu, K.; Gong, G.; and Mu, Y. 2020. Beyond Short-Term Snippet: Video Relation Detection with Spatio-Temporal Global Context. In *CVPR*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Liu, X.; Li, Y.-L.; and Lu, C. 2022. Highlighting Object Category Immunity for the Generalization of Human-Object Interaction Detection. In *AAAI 2022*.

Liu, X.; Li, Y.-L.; Wu, X.; Tai, Y.-W.; Lu, C.; and Tang, C.-K. 2022. Interactiveness Field in Human-Object Interactions. In *CVPR*.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*.

Materzynska, J.; Xiao, T.; Herzig, R.; Xu, H.; Wang, X.; and Darrell, T. 2020. Something-Else: Compositional Action Recognition with Spatial-Temporal Interaction Networks. In *CVPR*.

Ni, Z.; Valls Mascaró, E.; Ahn, H.; and Lee, D. 2023. Human–Object Interaction Prediction in Videos through Gaze Following. *Computer Vision and Image Understanding*, 233: 103741.

Prokudin, S.; Lassner, C.; and Romero, J. 2019. Efficient Learning on Point Clouds With Basis Point Sets. In *ECCV*.

Qi, S.; Wang, W.; Jia, B.; Shen, J.; and Zhu, S.-C. 2018. Learning human-object interactions by graph parsing neural networks. In *ECCV*.

Rajasegaran, J.; Pavlakos, G.; Kanazawa, A.; and Malik, J. 2022. Tracking People by Predicting 3D Appearance, Location & Pose. In *CVPR*.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *CVPR*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized Intersection over Union. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; and Tomasi, C. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*.

Sadeghian, A.; Alahi, A.; and Savarese, S. 2017. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *CVPR*.

Sadhu, A.; Chen, K.; and Nevatia, R. 2020. Video Object Grounding using Semantic Roles in Language Description. In *CVPR*.

Shan, D.; Geng, J.; Shu, M.; and Fouhey, D. F. 2020. Understanding Human Hands in Contact at Internet Scale. In *CVPR*.

Shang, X.; Ren, T.; Guo, J.; Zhang, H.; and Chua, T.-S. 2017. Video visual relation detection. In *ACMMM*.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*.

Wang, X.; and Gupta, A. 2018. Videos as space-time region graphs. In *ECCV*.

Weinzaepfel, P.; Martin, X.; and Schmid, C. 2016. Human action localization with sparse spatial supervision. *arXiv preprint arXiv:1605.05197*.

Wu, X.; Li, Y.-L.; Liu, X.; Zhang, J.; Wu, Y.; and Lu, C. 2022. Mining Cross-Person Cues for Body-Part Interactiveness Learning in HOI Detection. In *ECCV*.

Xu, X.; Li, Y.-L.; and Lu, C. 2022. Learning to Anticipate Future with Dynamic Context Removal. In *CVPR*.

Yang, Z.; Mahajan, D.; Ghadiyaram, D.; Nevatia, R.; and Ramanathan, V. 2019. Activity driven weakly supervised object detection. In *CVPR*.

Yao, L.; Han, J.; Wen, Y.; Liang, X.; Xu, D.; Zhang, W.; Li, Z.; Xu, C.; and Xu, H. 2022. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35: 9125–9138.

Yuan, Y.; Liang, X.; Wang, X.; Yeung, D.-Y.; and Gupta, A. 2017. Temporal dynamic graph LSTM for action-driven video object detection. In *ICCV*.

Zhang, F. Z.; Yuan, Y.; Campbell, D.; Zhong, Z.; and Gould, S. 2023. Exploring Predicate Visual Context in Detecting Human–Object Interactions. In *ICCV*.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; and Misra, I. 2022. Detecting Twenty-thousand Classes using Image-level Supervision. In *ECCV*.

Zhuo, T.; Cheng, Z.; Zhang, P.; Wong, Y.; and Kankanhalli, M. 2019. Explainable video action reasoning via prior knowledge and state transitions. In *ACMMM*.