HOI4ABOT: Human-Object Interaction Anticipation for Human Intention Reading Assistive roBOTs

Anonymous Author(s)
Affiliation
Address
email

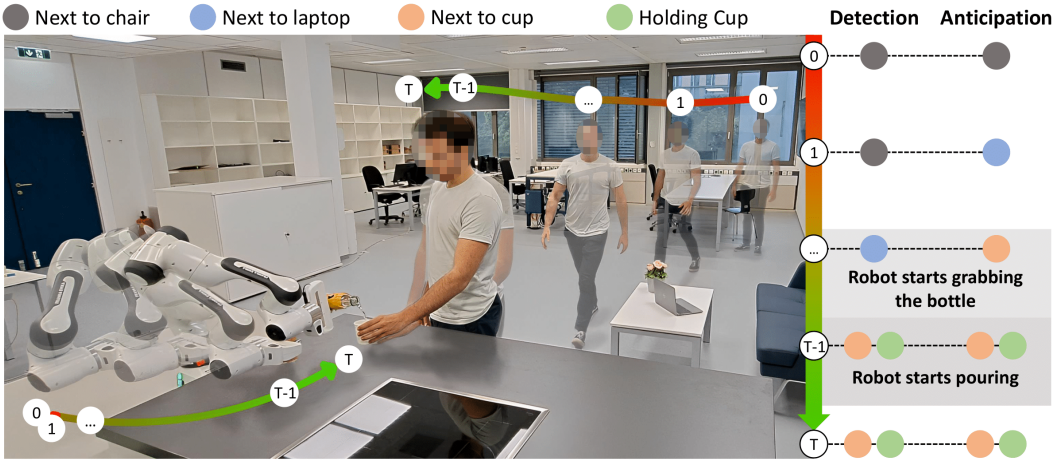


Figure 1: **Overview of our HOI4ABOT framework.** A robot leverages RGB data to detect and anticipate the human-object interactions in its surroundings and assist the human in a timely manner. The robot anticipates the human intention of holding the cup, so it prepares itself for pouring by grabbing the bottle. The robot reacts to the human holding the cup by pouring water.

Abstract: Robots are becoming increasingly integrated into our lives, assisting us in various tasks. To ensure effective collaboration between humans and robots, it is essential that they understand our intentions and anticipate our actions. In this paper, we propose a Human-Object Interaction (HOI) anticipation framework for Assistive Robots. We propose an efficient and robust transformer-based model to detect and anticipate HOIs from videos. This enhanced anticipation empowers robots to proactively assist humans, resulting in more efficient and intuitive collaborations. Our model outperforms state-of-the-art results in HOI detection and anticipation in VidHOI dataset with an increase of 1.76% and 1.04% in mAP respectively while being 15.4 times faster. We showcase the effectiveness of our approach through experimental results in a real robot, demonstrating that the robot's ability to anticipate HOIs is key for better Human-Robot Interaction.

Keywords: Human-Object Interaction, Assistive Robot, Human Intention

1 Introduction

In recent years, the field of robotics has witnessed significant interest in human-robot interaction (HRI), with a focus on enhancing the ability of robots to assist humans in various tasks [1, 2, 3, 4]. To facilitate effective collaboration between humans and robots, it is crucial for the robot to possess an understanding of both the surrounding environment and the individuals within it, including their

Submitted to the 7th Conference on Robot Learning (CoRL 2023). Do not distribute.

intentions. For example, consider the scenario visualized in Fig 1 where a robot is assisting a person in the kitchen. By recognizing the person's intention to prepare a drink and understanding their actions such as reaching for the cup, the robot can proactively provide the necessary support in a timely manner, such as picking up a bottle and pouring water. Therefore, by recognizing and anticipating human-object interactions (HOIs), the robot gets a solid understanding of the person's intention and better caters to their needs [1].

While HOI is a long-standing challenge in the computer vision community, most approaches only consider the detection of these interactions from single frames [5, 6, 7, 8, 9, 10]. However, to minimize the latency when a person is assisted by a robot, the detection is not enough, but anticipating is needed [11]. Therefore, we consider the task of HOI detection and anticipation, and we propose to leverage temporal cues from videos to better understand human intention. HOIs recognition in videos has been explored recently [12, 13, 14, 15]. In this paper, we propose an efficient deep learning architecture that combines pre-trained models with spatio-temporal consistency to successfully detect and anticipate HOIs in real-time. Our proposed model outperforms the state-of-the-art in VidHOI dataset [12] in terms of accuracy and speed. Moreover, we ensemble our framework with behavior trees [16] approach, that adapts in real-time the robot actions to enhance a better interaction with the human. We implement our framework in a real robot and demonstrate the effectiveness of our approach in a 'pouring task', showcasing the robot's ability to anticipate HOIs and proactively assist the human while reducing latency in the execution.

Our study presents a comprehensive framework for human-object interaction detection and anticipation, leveraging computer vision and deep learning techniques. By equipping robots with the ability to anticipate HOIs, we enable them to proactively assist humans, facilitating more efficient and intuitive collaborations. The contributions of our paper are summarized next:

1. A real-time transformer-based model that for HOIs detection and anticipation.
2. A novel patch merging strategy to align image features to pre-extracted bounding boxes.
3. The proposed HOI4ABOT outperforms the state-of-the-art in VidHOI across all HOI tasks.
4. To the best of our knowledge, we are the first to assess HOIs anticipation in a real robot experiment for an assistive task.

2 Related Works

2.1 Human Intention in Robotics

Recognizing and predicting human intention is crucial to ensure seamless collaboration between humans and robots [1, 17, 18]. Recent works in computer vision have highlighted the potential of harnessing human intention to improve the ability to anticipate future human actions [19, 20, 21]. These works employ a hierarchical structure in which human intention guides their model to predict future actions. In particular, [21] leverages the detection of human-object interactions (HOIs) within a scene to understand this high-level intention of the individuals. Despite the benefits of using HOIs to better understand human actions, their application in robotics from vision data has not been extensively explored [1]. [4] proposes a conditional random field (CRF) to assess the feasibility of a robot executing a given task based on the anticipated human actions. The CRF predicts the next human actions by considering the predicted object affordances and positions in the future. However, [4] is not scalable to new tasks as the CRF relies on hand-crafted features and is limited to a set of pre-determined affordance types. Instead, we train our model in the largest HOI video dataset available to learn robust features that enhance the robot's ability to anticipate human intention. Recently, [22] considers a spatial-attention network to extract scene graphs from images in an industrial scenario. However, [22] neglects the time-dependency in the task structure and proposes a reactive behavior that can not anticipate the human intention and enhance the robot's assistive ability. In this work, we propose a proactive robot behavior that assists a human based on the anticipation of human intention by predicting HOIs. Additional works on scene graphs only focus on the relationships of the objects for task planning [23, 24, 25].

2.2 HOI Detection and Anticipation

HOI focuses on localizing the humans and objects in a scene and classifying their interactions using a ⟨human, interaction, object⟩ triplet (e.g. ⟨person1, hold, cup⟩). HOI task has recently gained attention in the computer vision community due to its promising applications in downstream tasks, such as scene understanding [26] or action recognition [27]. The primary focus in the field revolves around the detection of HOI from images [5, 6, 7, 8, 9, 10]. Some [7, 8, 9] adopt a one-stage approach, directly operating on the images to predict the HOI triplet. However, these methods require higher training resources and do not benefit from pre-trained object detections. On the contrary, [5, 6, 10] employ a two-stage method to first locate the objects and humans in the image using pre-trained models and then classify each interaction using multi-stream classifiers. In particular, [10] uses a ViT transformer [28] to extract the patched features and proposes Masking with Overlapped Area (MOA) to extract features per object or human through a self-attention layer. Our work shows that weighting the patched features is sufficient to outperform MOA while not requiring any additional parameters.

While processing individual frames may be adequate for HOI detection, we argue that HOI anticipation benefits from leveraging the temporal aspects inherent in these interactions. Several studies in HOI detection address this temporal dimension by focusing on videos [12, 13, 14, 15]. [14] fuses patched features at multiple levels to generate instance representations utilizing a deformable tokenizer. [12] employs a two-stage model that uses 3D convolutions to merge features across the temporal dimension. [13] also adopts a two-stage approach but relies on a spatio-temporal transformer [29] to detect the interactions in videos. Finally, [15] extends the architecture from [13] by concatenating the human and object temporal features and fusing them with the human gaze information using cross-attention. [15] is the first work to propose both HOIs detection and anticipation in videos. Similarly to [10], [15] also adopts focal loss [30] to tackle the HOI imbalance in training. We adopt the findings from [15] but observe their model to not be feasible to work in real-time. Moreover, [15] trains a unique model for each anticipation horizon in the future. Instead, we propose a novel real-time multi-head model that can detect and anticipate HOIs in a single step.

2.3 Task and Motion Planning

For a robot to effectively assist and collaborate with a human in a particular task, it is essential to represent and understand the structure and order of actions involved, enabling the robot to achieve desired goals [31]. Finite State Machines (FSM) have long been the standard choice for control sequence in robotics [32, 33]. However, scaling FSM poses a challenge due to their lack of modularity and flexibility [16]. Recently, Behavior Trees (BT) [16] have gained popularity as they can facilitate task planning in human-robot collaborative tasks [34, 35], where the environment is dynamic. Our work adopts BT and defines its behavior based on the anticipate human intention and its uncertainty. Once a suitable chain of actions has been found by the task planner, motion planning is responsible for determining the low-level movements of the robot. Motion planning is a core problem in robotics [36, 37, 38, 39, 40]. [36, 37] propose to randomly sample points in the state space towards the goal. However, they consider humans as obstacles or constraints, not collaborators. Some approaches [38, 39] formulate motion planning as an optimization problem, but their applications in HRI are limited as determining the cost function related to humans is not trivial. Alternatively, motion generators can be learned from human demonstrations to obtain more natural movement [40, 41]. Dynamic Movement Primitives (DMPs) [40] have been successfully employed in HRI, by dynamically adapting the parameters [42, 43, 44].

3 Methodology

In this section, we present our **H**uman-**O**bject **I**nteraction **A**nticipation framework for **A**ssistive ro**BOT**s, called **HOI4ABOT**. First, we formulate the HOI detection and anticipation task. Then, we describe the integration of the deep learning architecture into the robot framework.

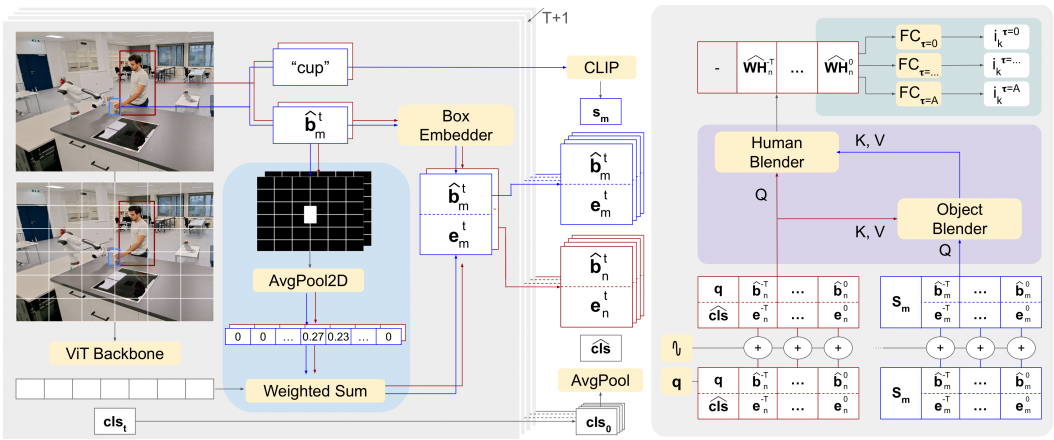


Figure 2: **HOI4ABOT architecture overview.** Our two-stage model predicts HOI in videos. Here, we consider a video of $T+1$ frames with the pre-extracted object and human bounding boxes. Our module initially extracts relevant features per frame (left) to later on detect and anticipate HOIs (right). First, we leverage a pre-trained ViT architecture [45] to extract patch-based local features $\mathbf{E}^t$ and also global context features $\mathbf{cls}_t$ per each frame. Then, we obtain individual features per human $\mathbf{e}_n^t$ and object $\mathbf{e}_m^t$ by aligning $\mathbf{E}^t$ to their bounding boxes through our Patch Merger strategy, shown in blue. We also project each bounding box $\mathbf{B}^t$ to $\hat{\mathbf{B}}^t$ using a box embedder [46], and the object category to a semantic feature using CLIP [47]. We finalize the feature extraction by averaging all $\mathbf{cls}_t$ in time. Our Dual Transformer module, shown in purple, leverages the human and object-constructed windows (sequences in red and blue respectively) through two cross-attention transformers. Finally, we project the enhanced last feature from the Human Blender to detect and anticipate HOIs at several time horizons i_k^τ in the future through our Hydra head (shown in green).

3.1 Human-Object Interaction

Let $\mathbf{V} = [\mathbf{f}_{-T}, \cdots, \mathbf{f}_0]$ be a frame sequence of duration T. The goal is to predict the interaction class i_k^τ in the subsequent time τ between any human $\mathbb{H}_n$ and object $\mathbb{O}_m$ pair $\mathbb{P}_k = \{\mathbb{H}_n, \mathbb{O}_m\}$ observed during the video $\mathbf{V}$, where $0 \leq n \leq N, 0 \leq m \leq M, 0 \leq k \leq K = M * N$. A visual illustration of our HOI4ABOT architecture is depicted in Fig.2.

Detection and tracking. HOI4ABOT is a two-stage method. First, we leverage off-the-shelf state-of-the-art object detection and tracking methods to identify the bounding boxes $\mathbf{B}_m \in \mathbb{R}^{T\times 4}$, label c_m, and track identifier id_m for any object $\mathbb{O}_m = \{id_m, c_m, \mathbf{B}_m\}$ in the video $\mathbf{V}$. $\mathbf{B}_m = [\mathbf{b}_m^{-T}, \cdots, \mathbf{b}_m^0]$ represents a list of XY pixel coordinates of the top-left corner and right-bottom corner of the bounding box that locates a given object $\mathbb{O}_m$ at each frame $\mathbf{f}_t$ of the observed video. We obtain the same information for each human $\mathbb{H}_n$. In the second stage, we exploit each individual pair $\mathbb{P}_k = \{\mathbb{H}_n, \mathbb{O}_m\}$ to predict its interaction class i_k^τ in a given time horizon τ using various data modalities. This requires understanding the visual features $\mathbf{V}$ of the pair, how their spatial relationship evolves through time $\mathbf{B}_k = [\mathbf{B}_n, \mathbf{B}_m]$ and also the intrinsic semantics of the object c_m.

Visual features. We use Dinov2 [45] as a pre-trained Visual Transformer (ViT) [28] backbone to divide each frame $\mathbf{f}_t$ into $L \times L$ patches and project each patch $\mathbf{p}_l^t$ to a visual token $\mathbf{e}_l^t$ that encodes the image information of that patch l. In total, the image encoder obtains $\mathbf{E}^t \in \mathbb{R}^{L^2\times d}$ that captures the local visual features, plus the global context vector $\mathbf{cls}_t \in \mathbb{R}^d$ of a frame $\mathbf{f}_t$.

We develop a simple but efficient technique to extract individual features per human and object from a frame through a single step. We call this method Patch Merger. For simplification, let us consider an object $\mathbb{O}_m^t$ and its box $\mathbf{b}_m^t$ at frame $\mathbf{f}_t$. First, we create a binary mask with the same dimensions as $\mathbf{f}_t$, where 1 denotes a pixel laying within $\mathbf{b}_m^t$. We convert the binary mask in a sequence of patches following [28]. Then, we obtain a weighting vector $\boldsymbol{\omega}_m^t$ by computing the percentage that $\mathbf{b}_m^t$ overlaps each patch through appropriate 2D Average Pooling and normalization. Finally, we compute the weighted sum of local visual features $\mathbf{e}_m^t = \sum \boldsymbol{\omega}_m^t \mathbf{E}^t$, obtaining the individual representation of

$\mathbb{O}_m^t$. Compared to [10], which normalizes along the patch dimension and uses a quantized sequence as the attention mask for a self-attention layer, our algorithm is parameter-free, more efficient, and shows better performance in our experiments.

While relevant works [13, 15] extract the spatial relationships of the objects and humans within each frame using an attention mechanism [29], we consider the class token $\mathbf{cls}_t \in \mathbb{R}^d$ to already encode the context of the scene. We consider the context to be homogeneous among frames, thus we summarize all $\mathbf{cls}_t$ into a global vector $\widehat{\mathbf{cls}} \in \mathbb{R}^d$ per video though average pooling.

Spatial features. For each bounding box $\mathbf{b}_m^t$, we extract the XY normalized pixel coordinates for the top-left corner and right-bottom corner. Then, we adopt a positional encoding using random spatial frequencies [46] to embed the location of each point and merge these two corner representations into one box representation $\hat{\mathbf{b}}_m^t \in \mathbb{R}^d$ using a fully connected layer. This process is also applied to humans, thus obtaining $\hat{\mathbf{b}}_n^t \in \mathbb{R}^d$ to encode each human $\mathbb{H}_n^t$ position in the scene.

Object semantics. Leveraging the object semantics is key to understanding the possible interactions in a given pair. While 'holding a cup' or 'holding a bottle' are both feasible, 'holding a car' becomes more unrealistic. Therefore, we extract object semantic information $\mathbf{s}_m \in \mathbb{R}^d$ per object $\mathbb{O}_m$ to facilitate the model predicts the intention class i_k^τ. For that, we use the CLIP text encoder [47].

Pair Interaction. Given all the previously extracted features, we construct a temporal architecture that leverages the evolution of the interactions between a human $\mathbb{H}_n$ and an object $\mathbb{O}_m$ in time. We process each pair independently, and therefore we focus on a single human-object pair in the formulation. We construct a sequence of tokens for each entity of the pair independently. For any entity, we stack both the visual tokens $\mathbf{E}_n = [\mathbf{e}_n^{-T}, \cdots, \mathbf{e}_n^0]$ and the spatial features $\hat{\mathbf{B}}_n = [\hat{\mathbf{b}}_n^{-T}, \cdots, \hat{\mathbf{b}}_n^0]$ in time. We then obtain a human temporal window $\mathbf{WH}_n = [\hat{\mathbf{B}}_n, \mathbf{E}_n]$ and the object's temporal window $\mathbf{WO}_m = [\hat{\mathbf{B}}_m, \mathbf{E}_m]$ which are constructed with frame-wise local information each. We sum to both $\mathbf{WH}_n$ and $\mathbf{WO}_m$ a sinusoidal positional encoding to inform the attention mechanism about the sequence order. Then, to provide more context, we extend $\mathbf{WH}_n$ by prepending the global visual feature and a learnable spatial parameter $[\mathbf{q}, \widehat{\mathbf{cls}}]$. $\mathbf{q}$ learns the evolution of the location of the human in time through the attention mechanism. We also extend $\mathbf{WO}_m$ by prepending the semantic token $\mathbf{s}_m$ that encodes the object type. Therefore, we obtain a temporal feature $\mathbf{WH}_n \in \mathbb{R}^{(T+2)\times d}$ and $\mathbf{WO}_m \in \mathbb{R}^{(T+2)\times d}$ per pair.

To extract the HOI relationships between $\mathbb{H}_n$ and $\mathbb{O}_m$, we train a dual transformer with cross-attention layers. First, an Object Blender transformer enhances the object window $\mathbf{WO}_m$ based on the human knowledge $\mathbf{WH}_n$. Then, the blended object features $\widehat{\mathbf{WO}}_m$ are used to extend the human representation $\mathbf{WH}_n$ in the Human Blender transformer to $\widehat{\mathbf{WH}}_n$. Finally, we extract the last token from $\widehat{\mathbf{WH}}_n$, which encodes the most current status of the scene, and classify the interaction pair i_k^τ using a fully connected layer. As a given human-object pair can have multiple interactions simultaneously, we use a sigmoid function and define a threshold to classify the current interactions.

Multi-head for multiple future horizons. The goal is to predict the interaction class i_k^τ in the subsequent time τ between any human $\mathbb{H}_n$ and object $\mathbb{O}_m$ pair $\mathbb{P}_k = \{\mathbb{H}_n, \mathbb{O}_m\}$. We considered the problem of HOI detection ($\tau = 0$) and also the anticipation in multiple future horizons ($\tau > 0$). Contrary to [15] that proposes one trained model for each τ, we developed a single model that can predict multiple time horizon interactions. For that, we froze the HOI4ABOT trained in the detection task, and train an additional linear layer that projects the last token from $\widehat{\mathbf{WH}}_n$ to the interaction for the particular τ. We call this shared backbone the Hydra variant, which allows us to simultaneously predict interactions across multiple τ, making our model faster and more efficient. We consider our Hydra variant with A number of heads.

3.2 Motion generation and task planning

Motion Generation. The proposed framework segments the complex movements into simpler movement primitives, which are learned with DMPs. To collect demonstrations of each move-

ment primitive, we employ kinesthetic teaching, where an operator guides the robot's end effector by physically manipulating it [48]. Generating the motion requires estimating the goal position, which we obtain through the use of a calibrated vision system that relies on a pre-trained object detector (i.e. YOLOv8 [49]) and a depth camera. The position of the goals with respect to the base of the robot is computed using the intrinsic and extrinsic camera matrices.

Task planning. Properly scheduling the acquired movement primitives is crucial to reach a desired goal. We implement Behavior Trees (BT) [16] as a ROS node that subscribes to the predicted HOIs and their confidence. The reactiveness of BTs allows adapting the robot's behavior by considering the anticipated human intention and changing to the appropriate sub-tree if needed. This is motivated by how humans interact with each other. For example, if a bartender observes a client approaching the bar, they can prepare for the interaction by grabbing a glass, thus reducing the serving time.

Robot control. The generated poses from the motion generator are passed to the controller. In our system, we employ a Cartesian impedance controller [50, 51] to achieve the compliant behavior of the manipulator. This controller enhances the safety of human-robot collaboration by allowing the robot to respond in a compliant manner to external forces and disturbances.

4 Experiments

In this section, we evaluate the effectiveness of our HOI4ABOT both quantitatively and qualitatively. First, we introduce the benchmarks, metrics, and results of our stand-alone deep learning architecture. Then, we conduct an ablation study to reinforce architectural decisions. Finally, we present our qualitative results in a real experiment.

4.1 Dataset and Metrics

We train and evaluate our model on the VidHOI dataset [12], the largest dataset available for human-object interactions in videos. This dataset encompasses 7.3 million frames with 755,000 annotated interactions of one frame per second. To assess the performance of our approach, we adopted the same evaluation metrics as those presented in [15]. We computed the mean average precision (mAP) using the method presented in [52]. The mAP@50 incorporates the precision-recall curves for all interaction classes. To determine a correct HOI triplet, three conditions need to be met: (i) the detected bounding boxes for human and object must overlap with their corresponding ground truths with an Intersection over Union (IoU) of 50 %, (ii) the predicted object category is correct, (iii) the predicted interaction is correct. Following standard evaluation in VidHOI, we report mAP across three different HOI sets: (i) Full: all interaction categories, (ii) NonRare: frequent interactions in the validation set (more than 25 appearances), (iii) Rare: non-frequent interactions (less than 25). Additionally, we also evaluated our approach in two different modes: Oracle and Detection. In the Oracle mode, we utilized ground truth detections, whereas, in the Detection mode, we employed the YOLOv5 [53] object detector. To ensure a fair comparison, we employed the identical weights for the object detector as those specified in [15]. Furthermore, we computed the Person wise top-k metrics as outlined in [15]. In these metrics, a detection was considered correct if one of the top-k predicted interactions matched the ground truth. We computed this metric for each person in the validation set and averaged the results over all humans.

4.2 Quantitative evaluation

HOI4ABOT outperforms state-of-the-art models in terms of accuracy and speed across all different tasks and scenarios. Table 1 shows that our HOI4ABOT outperforms the state-of-the-art [12, 52, 14, 13, 15] across all different sets in the oracle and detection mode. Moreover, Table 2 shows how efficiently reusing our pre-trained strong backbone for detection helps our Hydra variant to outperform all models in the anticipation task for all reported metrics. Table 2 also shows that using the proposed pre-training strategy outperforms training from scratch with a separate model for each anticipation horizon.

Table 1: Detection mAP.

Method	Full	Non-Rare	Rare
Oracle Mode			
ST-HOI [12]	17.6	27.2	17.3
QPIC [52]	21.4	32.9	20.56
TUTOR [14]	26.92	37.12	23.49
STTran [13]	28.32	42.08	17.74
ST-Gaze [15]	38.61	52.44	27.99
Ours (Dual)	40.37	**54.52**	29.5
Ours (Stacked)	**40.55**	53.94	**30.26**
Detection Mode			
STTran [13]	7.61	13.18	3.33
ST-Gaze [15]	10.4	16.83	5.46
Ours (Dual)	**11.12**	**18.48**	**5.61**
Ours (Stacked)	10.79	17.79	5.42

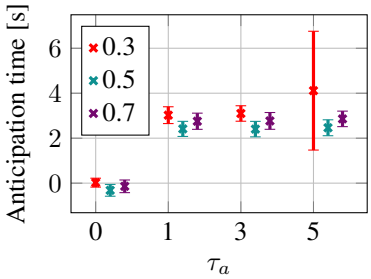
Table 2: Anticipation mAP in Oracle mode.

Method	τ_a	mAP	Preson-wise top-5			
			Rec	Prec	Acc	F1
STTran [13]	1	29.09	**74.76**	41.36	36.61	50.48
	3	27.59	**74.79**	40.86	36.42	50.16
	5	27.32	**75.65**	41.18	36.92	50.66
ST-Gaze [15]	1	37.59	72.17	59.98	51.65	62.78
	3	33.14	71.88	60.44	52.08	62.87
	5	32.75	71.25	59.09	51.14	61.92
Dual (scratch)	1	**38.46**	73.32	63.78	55.37	65.59
	3	34.58	73.61	61.7	54	64.48
	5	33.79	72.33	63.96	55.28	65.21
Dual (HYDRA)	1	37.77	74.07	**64.9**	**56.38**	**66.53**
	3	**34.75**	74.37	**64.52**	**56.22**	**66.4**
	5	**34.07**	73.67	**65.1**	**56.31**	**66.4**

Table 3: Detection mAP for different model variants.

Variant	mAP
Feature blender = MOA	40
Interaction token = Learnable	40.29
Main branch = object	39.85
Transformer type = single	40.26
Transformer type = stacked	**40.55**
Dual	40.37

Figure 3: Time in advance to anticipate a HOIs and its confidence thresholds.



4.3 Ablation study

This section analyses our proposed approaches and their impact on the performance of the HOI task. All results are depicted in Table 3. For simplification, we only consider the HOI detection task.

Firstly we explore different variations in the extraction and arrangement of features to compose the human and object windows. We compare our Patch Merger strategy to the MOA strategy reported in [10]. Using MOA requires an additional self-attention block, which increases the model's parameters while underperforming. Moreover, we explore different feature aggregation strategies to classify an interaction. Instead of using the last observed token in $\widehat{\mathbf{WH}}_n$ for classification, we prepend an additional learnable token to $\mathbf{WH}_n$ which aggregates the interaction relationships, inspired by the ViT class token [28]. However, Table 3 shows that classifying from the last observed features is better while not requiring additional parameters. Last, we consider varying the order of the cross-attentions branches, first the Human Blender and second the Object Blender. We consider the reason it underperforms is the conceptual difference in behavior between humans and objects: objects are static and therefore less informative than humans, which are dynamic and lead the interaction.

Secondly, we assess our dual transformer by comparing it with other variants. We consider the 'single' variant when only using the Human Blender transformer, which underperforms as not being able to effectively capture the HOIs. We also consider stacking both $\mathbf{WH}_n \in \mathbb{R}^{(T+2)\times d}$ and $\mathbf{WO}_m \in \mathbb{R}^{(T+2)\times d}$ to a single feature window pair, $\mathbf{WP}_k \in \mathbb{R}^{(T+2)\times 2d}$. We observe slight improvements in this variant in terms of mAP when detecting in the Oracle mode, but it underperforms in the detection mode and for the anticipation tasks, as shown in the supplementary materials.

Finally, we compare the inference time of our model to [15] to assess the efficiency in real-world applications in robots. Our 'dual' variant is 15.4 times faster than [15] for the detection task. [15] requires extracting gaze maps, which drastically slows down the inference speed of their model. When using our Hydra model, we obtain interactions for the time horizons 0, 1, 3, and 5 using one single forward pass, with nearly the same inference speed and parameters as using one single head. More information can be found in the supplementary materials.

4.4 Real World Experiments

HOI detection and anticipation are essential for robots to comprehend the surrounding humans and better predict their needs, so the robot can assist in a timely manner. We conduct real experiments with a Franka Emika Panda robot to showcase the benefit of our approach in assistive robots beyond the offline VidHOI dataset. The VidHOI dataset contains user-collected videos of humans, mostly performing outdoor activities, and that can not be easily related to robotic assistive tasks. We consider the 'pouring task' in a kitchen scenario where the robot assumes the role of a bartender with the goal of pouring a beverage for the human. The scenario is shown in Fig. 1 and is the only one appropriate for assistive robots based on the HOIs annotated in the dataset. To assess the performance of our model in unseen scenarios, we collect 20 videos in our kitchen lab. The human is instructed to grab the cup and informed that will be assisted by the robot in the pouring. We manually annotate the time the person grabs the cup to use as ground truth.

Our Hydra variant detects and anticipates the HOI between a person and a cup in real time. When the robot anticipates that the human will be near the cup, it proceeds to grab the bottle. However, if the human moves away the robot releases the bottle and returns to the initial pose. The robot proceeds to pour the liquid into the cup after detecting that the human is holding it. We evaluate how much time in advance our model anticipates HOIs for different heads of our Hydra variant and confidence thresholds. Our results, shown in Fig. 3, demonstrate that HOI anticipation facilitates earlier prediction of the human intention, thus reducing the robot latency in the pouring task. We implement and evaluated [15] for the detection task and observe lag in the model prediction despite only assessing two interactions and one τ_a, reinforcing the need for an efficient model and the benefit of our approach. Additional qualitative results are provided in the supplementary video.

5 Limitations

Despite outperforming state-of-the-art models in HOI from videos, in both detection and anticipation tasks, we observe from qualitative experiments the challenge of the implementation in the real world. Our model performs with higher confidence for spatial interaction classes, such as a human being 'next to' or 'in front of' an object. However, despite 'holding a cup' being correctly detected, the anticipation of this interaction is challenging. We explore the VidHOI dataset and observe that most people already appear with the cup in their hand during the whole video. Therefore, the model struggles in anticipating this grabbing process as the behavior is not frequent in our dataset, despite being the largest one available for HOI in videos. To overcome this issue, we sample more frequently clips where the interaction changes in the anticipation horizon. Still, this is insufficient to ensure correct anticipation in 'holding a cup' with higher confidence. Future research directions consider training with a dataset more coupled to our robotics scenario to improve the model predictions. This would allow us to extend our experiments to more complex daily scenarios. Finally, the employed DMPs could be expanded or replaced by visual servoing to consider goal-following behaviors.

6 Conclusions

In this paper, we proposed a **H**uman-**O**bject **I**nteraction **A**nticipation for **A**ssistive ro**BOT**s framework (**HOI4ABOT**). We consider the task of detecting and anticipating human-object interactions (HOI) in videos through a transformer architecture. We train and evaluate HOI4ABOT in the VidHOI dataset and outperform current state-of-the-art across all tasks and metrics while being $15.4\times$ faster. Moreover, our model runs in real-time thanks to our efficient design. Additionally, we extend our HOI4ABOT model with a multi-head architecture, which can detect and anticipates HOIs across different future horizons in a single step. We demonstrate the effectiveness of our approach by implementing our model in a Franka Emika Panda robot. We show that anticipating HOIs in real-time is key for a robot to better assist a human in a timely manner and we support our findings with real experiments. In conclusion, our approach demonstrates its effectiveness and defines a new road to explore, where vision serves a leading role for assistive robots in collaboration scenarios.

References

[1] N. Robinson, B. Tidd, D. Campbell, D. Kulić, and P. Corke. Robotic vision for human-robot interaction and collaboration: A survey and systematic review. *J. Hum.-Robot Interact.*, 12(1), feb 2023. doi:10.1145/3570731. URL https://doi.org/10.1145/3570731.

[2] B. Abbasi, N. Monaikul, Z. Rysbek, B. D. Eugenio, and M. Žefran. A multimodal human-robot interaction manager for assistive robots. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6756–6762, 2019. doi:10.1109/IROS40897.2019.8968505.

[3] V. Ortenzi, A. Cosgun, T. Pardi, W. P. Chan, E. Croft, and D. Kulić. Object handovers: A review for robotics. *IEEE Transactions on Robotics*, 37(6):1855–1873, 2021. doi:10.1109/TRO.2021.3075365.

[4] H. S. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):14–29, 2016. doi:10.1109/TPAMI.2015.2430335.

[5] B. Xu, Y. Wong, J. Li, Q. Zhao, and M. S. Kankanhalli. Learning to detect human-object interactions with knowledge. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2019–2028, 2019. doi:10.1109/CVPR.2019.00212.

[6] O. Ulutan, A. S. M. Iftekhar, and B. Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. pages 13614–13623, 06 2020. doi:10.1109/CVPR42600.2020.01363.

[7] J. Lim, V. M. Baskaran, J. M.-Y. Lim, K. Wong, J. See, and M. Tistarelli. Ernet: An efficient and reliable human-object interaction detection network. *IEEE Transactions on Image Processing*, 32:964–979, 2023. doi:10.1109/TIP.2022.3231528.

[8] X. Wu, Y.-L. Li, X. Liu, J. Zhang, Y. Wu, and C. Lu. Mining cross-person cues for body-part interactiveness learning in hoi detection. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision – ECCV 2022*, pages 121–136, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19772-7.

[9] Y. Liao, A. Zhang, M. Lu, Y. Wang, X. Li, and S. Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20091–20100, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society. doi:10.1109/CVPR52688.2022.01949. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01949.

[10] J. Park, J.-W. Park, and J.-S. Lee. Viplo: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17152–17162, 2023.

[11] L. Psarakis, D. Nathanael, and N. Marmaras. Fostering short-term human anticipatory behavior in human-robot collaboration. *International Journal of Industrial Ergonomics*, 87:103241, 2022. ISSN 0169-8141. doi:https://doi.org/10.1016/j.ergon.2021.103241. URL https://www.sciencedirect.com/science/article/pii/S0169814121001591.

[12] M.-J. Chiou, C.-Y. Liao, L.-W. Wang, R. Zimmermann, and J. 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*, ICDAR '21, page 9–17, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385299. doi:10.1145/3463944.3469097. URL https://doi.org/10.1145/3463944.3469097.

[13] Y. Cong, W. Liao, H. Ackermann, B. Rosenhahn, and M. Yang. Spatial-temporal transformer for dynamic scene graph generation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16352–16362, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society. doi:10.1109/ICCV48922.2021.01606. URL https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.01606.

[14] D. Tu, W. Sun, xiongkuo min, G. Zhai, and W. Shen. Video-based human-object interaction detection from tubelet tokens. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=kADW_LsENM.

[15] Z. Ni, E. Valls Mascaró, H. Ahn, and D. Lee. Human-object interaction prediction in videos through gaze following. *Computer Vision and Image Understanding*, page 103741, 2023. ISSN 1077-3142. doi:https://doi.org/10.1016/j.cviu.2023.103741. URL https://www.sciencedirect.com/science/article/pii/S1077314223001212.

[16] M. Colledanchise and P. Ogren. *Behavior Trees in Robotics and AI: An Introduction*. 07 2018. ISBN 9781138593732. doi:10.1201/9780429489105.

[17] Z. Li, Y. Mu, Z. Sun, S. Song, J. Su, and J. Zhang. Intention understanding in human–robot interaction based on visual-nlp semantics. *Frontiers in Neurorobotics*, 14:610139, 2021.

[18] F. Semeraro, A. Griffiths, and A. Cangelosi. Human–robot collaboration and machine learning: A systematic review of recent research. *Robotics and Computer-Integrated Manufacturing*, 79:102432, 2023. ISSN 0736-5845. doi:https://doi.org/10.1016/j.rcim.2022.102432. URL https://www.sciencedirect.com/science/article/pii/S0736584522001156.

[19] E. V. Mascaro, H. Ahn, and D. Lee. Intention-conditioned long-term human egocentric action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6048–6057, January 2023.

[20] D. Roy and B. Fernando. Action anticipation using latent goal learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2745–2753, January 2022.

[21] P. Wei, Y. Liu, T. Shu, N. Zheng, and S.-C. Zhu. Where and why are they looking? jointly inferring human attention and intentions in complex tasks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6801–6809, 2018. doi:10.1109/CVPR.2018.00711.

[22] S. Li, P. Zheng, Z. Wang, J. Fan, and L. Wang. Dynamic scene graph for mutual-cognition generation in proactive human-robot collaboration. *Procedia CIRP*, 107:943–948, 2022. ISSN 2212-8271. doi:https://doi.org/10.1016/j.procir.2022.05.089. URL https://www.sciencedirect.com/science/article/pii/S2212827122003730. Leading manufacturing systems transformation – Proceedings of the 55th CIRP Conference on Manufacturing Systems 2022.

[23] S. Amiri, K. Chandan, and S. Zhang. Reasoning with scene graphs for robot planning under partial observability. *IEEE Robotics and Automation Letters*, 7(2):5560–5567, 2022.

[24] C. Agia, K. Jatavallabhula, M. Khodeir, O. Miksik, V. Vineet, M. Mukadam, L. Paull, and F. Shkurti. Taskography: Evaluating robot task planning over large 3d scene graphs. In *Conference on Robot Learning*, pages 46–58. PMLR, 2022.

[25] M. Patel and S. Chernova. Proactive robot assistance via spatio-temporal object modeling. In *6th Annual Conference on Robot Learning*, 2022.

[26] Y. Xu, J. Zhang, Q. Zhang, and D. Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022.

[27] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*, 2021.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems (NeurIPS)*, 30, 2017.

[30] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

[31] H. Guo, F. Wu, Y. Qin, R. Li, K. Li, and K. Li. Recent trends in task and motion planning for robotics: A survey. *ACM Comput. Surv.*, feb 2023. ISSN 0360-0300. doi:10.1145/3583136. URL `https://doi.org/10.1145/3583136`. Just Accepted.

[32] M. Awais and D. Henrich. Online intention learning for human-robot interaction by scene observation. In *2012 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, pages 13–18, 2012. doi:10.1109/ARSO.2012.6213391.

[33] I. El Makrini, M. Omidi, F. Fusaro, E. Lamon, A. Ajoudani, and B. Vandcrborght. A hierarchical finite-state machine-based task allocation framework for human-robot collaborative assembly tasks. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10238–10244, 2022. doi:10.1109/IROS47612.2022.9981618.

[34] C. Paxton, A. Hundt, F. Jonathan, K. Guerin, and G. D. Hager. Costar: Instructing collaborative robots with behavior trees and vision. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 564–571. IEEE, 2017.

[35] M. Carvalho, J. Avelino, A. Bernardino, R. Ventura, and P. Moreno. Human-robot greeting: tracking human greeting mental states and acting accordingly. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1935–1941, 2021. doi:10.1109/IROS51168.2021.9635894.

[36] L. Kavraki, P. Svestka, J.-C. Latombe, and M. Overmars. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Transactions on Robotics and Automation*, 12(4):566–580, 1996. doi:10.1109/70.508439.

[37] J. Kuffner and S. LaValle. Rrt-connect: An efficient approach to single-query path planning. In *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065)*, volume 2, pages 995–1001 vol.2, 2000. doi:10.1109/ROBOT.2000.844730.

[38] M. Kalakrishnan, S. Chitta, E. Theodorou, P. Pastor, and S. Schaal. Stomp: Stochastic trajectory optimization for motion planning. In *2011 IEEE International Conference on Robotics and Automation*, pages 4569–4574, 2011. doi:10.1109/ICRA.2011.5980280.

[39] C. Park, J. Pan, and D. Manocha. Itomp: Incremental trajectory optimization for real-time replanning in dynamic environments. 06 2012.

[40] P. Pastor, M. Kalakrishnan, F. Meier, F. Stulp, J. Buchli, E. Theodorou, and S. Schaal. From dynamic movement primitives to associative skill memories. *Robotics and Autonomous Systems*, 61:351–361, 04 2013. doi:10.1016/j.robot.2012.09.017.

[41] S. Calinon and D. Lee. *Learning Control*, pages 1261–1312. 01 2019. ISBN 978-94-007-6045-5. doi:10.1007/978-94-007-6046-2_68.

[42] M. Wu, B. Taetz, Y. He, G. Bleser, and S. Liu. An adaptive learning and control framework based on dynamic movement primitives with application to human–robot handovers. *Robotics and Autonomous Systems*, 148:103935, 2022. ISSN 0921-8890. doi:https://doi.org/10.1016/j.robot.2021.103935. URL https://www.sciencedirect.com/science/article/pii/S0921889021002189.

[43] M. Prada, A. Remazeilles, A. Koene, and S. Endo. Dynamic movement primitives for human-robot interaction: Comparison with human behavioral observation. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1168–1175, 2013. doi:10.1109/IROS.2013.6696498.

[44] M. Prada, A. Remazeilles, A. Koene, and S. Endo. Implementation and experimental validation of dynamic movement primitives for object handover. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2146–2153, 2014. doi:10.1109/IROS.2014.6942851.

[45] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[46] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020.

[47] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[48] M. Ito, K. Noda, Y. Hoshino, and J. Tani. Dynamic and interactive generation of object handling behaviors by a small humanoid robot using a dynamic neural network model. *Neural Networks*, 19(3):323–337, 2006. ISSN 0893-6080. doi:https://doi.org/10.1016/j.neunet.2006.02.007. URL https://www.sciencedirect.com/science/article/pii/S0893608006000311. The Brain Mechanisms of Imitation Learning.

[49] G. Jocher, A. Chaurasia, and J. Qiu. YOLO by Ultralytics, Jan. 2023. URL https://github.com/ultralytics/ultralytics.

[50] N. Hogan. Impedance control: An approach to manipulation. In *1984 American Control Conference*, pages 304–313, 1984. doi:10.23919/ACC.1984.4788393.

[51] N. Hogan. Impedance Control: An Approach to Manipulation: Part II—Implementation. *Journal of Dynamic Systems, Measurement, and Control*, 107(1):8–16, 03 1985. ISSN 0022-0434. doi:10.1115/1.3140713. URL https://doi.org/10.1115/1.3140713.

[52] M. Tamura, H. Ohashi, and T. Yoshinaga. QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, 2021.

[53] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, K. Michael, TaoXie, J. Fang, imyhxy, Lorna, Z. Yifu, C. Wong, A. V, D. Montes, Z. Wang, C. Fati, J. Nadar, Laughing, UnglvKitDe, V. Sonck, tkianai, yxNONG, P. Skalski, A. Hogan, D. Nair, M. Strobel, and M. Jain. ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation, Nov. 2022. URL https://doi.org/10.5281/zenodo.7347926.