Seeing My Future: Predicting Situated Interaction Behavior in Virtual Reality

Yuan Xu[1,*] Zimu Zhang[1,*] Xiaoxuan Ma[1] Wentao Zhu[2] Yu Qiao[3] Yizhou Wang[1]

[1] Peking University [2] Eastern Institute of Technology, Ningbo [3] Shanghai Jiao Tong University

Abstract

Virtual and augmented reality systems increasingly demand intelligent adaptation to user behaviors for enhanced interaction experiences. Achieving this requires accurately understanding human intentions and predicting future situated behaviors—such as gaze direction and object interactions—which is vital for creating responsive VR/AR environments and applications like personalized assistants. However, accurate behavioral prediction demands modeling the underlying cognitive processes that drive human-environment interactions. In this work, we introduce a hierarchical, intention-aware framework that models human intentions and predicts detailed situated behaviors by leveraging cognitive mechanisms. Given historical human dynamics and the observation of scene contexts, our framework first identifies potential interaction targets and forecasts fine-grained future behaviors. We propose a dynamic Graph Convolutional Network (GCN) to effectively capture human-environment relationships. Extensive experiments on challenging real-world benchmarks and live VR environment demonstrate the effectiveness of our approach, achieving superior performance across all metrics and enabling practical applications for proactive VR systems that anticipate user behaviors and adapt virtual environments accordingly.

1. Introduction

Anticipating and responding to user interactions represents the next frontier for creating more immersive and effective AR/VR experiences. Moving beyond reactive approaches, proactive adaptive systems promise to revolutionize applications spanning personalized gaming, training simulations, collaborative workspaces, and assistive technologies. To realize these systems, accurately predicting situated behaviors—where action arises from the interplay of intention and context—remains essential for enabling intelligent, adaptive systems across diverse applications.

For example, in virtual environments, predicting situated behavior enables VR systems or immersive games to adapt environment complexity or interaction modes in response to user intent, creating more engaging and personalized experiences [41, 58, 61]. A VR game might, for instance, predict which area the player is about to explore and introduce ambient story elements or adjust enemies accordingly. In the physical world, such predictive capabilities allow robots and AR systems to interpret user intent in context, offering proactive support for smoother human–AI collaboration [6, 33, 39].

Achieving such accurate behavioral prediction requires a deeper understanding of the cognitive processes underlying human–environment interactions. Human interaction behaviors are not random but emerge from underlying intentions that are continuously shaped and reshaped by dynamic situational contexts, such as environmental cues, task demands, and temporal constraints [1, 55–57]. Research in psychology and cognitive science [8, 9, 40, 43] suggests that humans often first form an interaction intention (*e.g.*, identifying potential objects for interaction), which then guides the planning and execution of specific actions. In the formation of interaction intention, the relationship between gaze and the environment plays a crucial role [32]. Gaze, by enabling the active search for relevant information in the environment, serves as the most direct and observable expression of attention [5]. It often precedes and guides interactions by locking onto targets that align with an individual's goals [24, 32, 46].

In light of this, we propose an intention-aware framework to predict comprehensive human situated behavior in an environment, including where they will look (*gaze*), where they will go (*trajectory*), and which objects they will interact with (*object interaction*), as shown in Fig. 1. The core of the proposed framework lies in a hierarchical modeling approach, inspired by the human cognitive process. Initially, it predicts a coarse set of potential interaction targets, such as a ketchup bottle or cake on a table based on historical gaze attention and human-scene interactions (see Fig. 1). From these, it identifies the top-K objects with the highest

* Equal contribution.

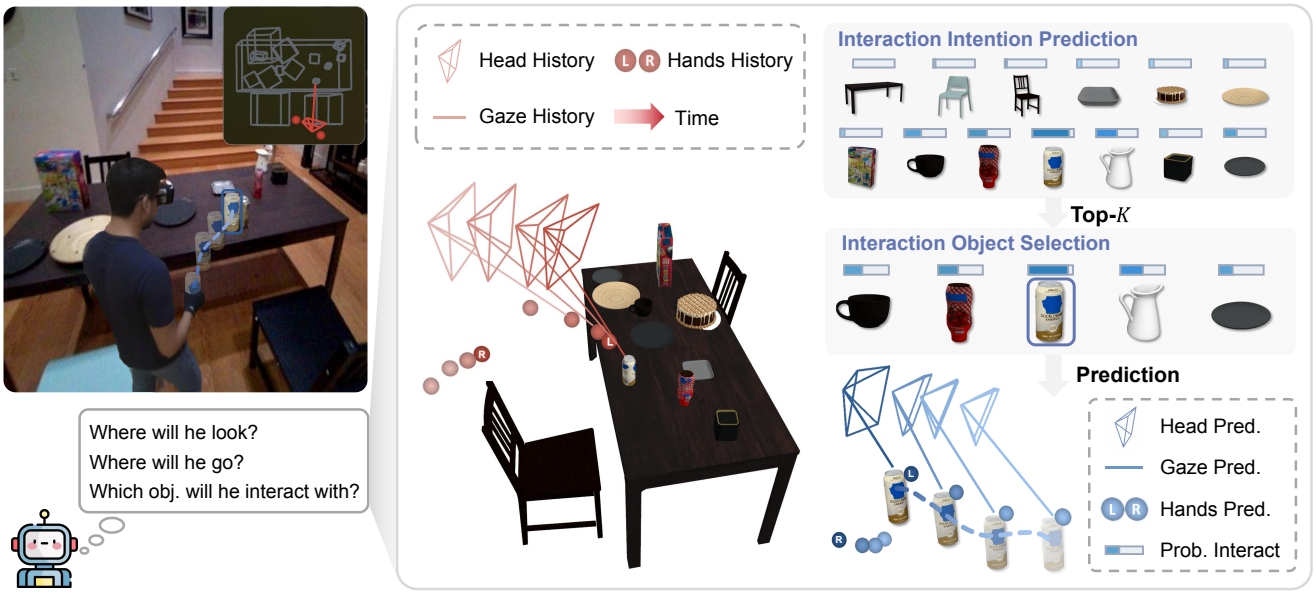


Figure 1. **Situated interaction behavior prediction**. Given historical human dynamics, such as gaze direction and situational context, we propose an intention-aware framework to predict a person's future behavior in the environment, including where to look (*gaze*), where to go (*trajectory*), and which objects to interact with (*object interaction*). Our approach employs a hierarchical prediction strategy aligned with human cognition, first coarsely identifying potential objects based on prior engagements before precisely forecasting the next specific interaction.

interaction likelihood. Subsequently, the framework forecasts detailed future behaviors, specifying both the object of interaction (*e.g.*, a coffee can as the final target) and the corresponding action, such as picking up the coffee can, as illustrated in Fig. 1. This predictive process is powered by a dynamic Graph Convolutional Network (GCN) to effectively capture the human-environment interactions and enhance the accuracy and robustness of behavior prediction in dynamic immersive environments.

To validate the practical applicability of our framework, we conduct user studies in a VR environment where participants engage in natural interaction sequences. Our results demonstrate robust performance even under noisy, realistic conditions, showing that the system can successfully anticipate user gaze patterns and future interaction behaviors, and highlighting the framework's potential to support a new generation of proactive VR applications.

To summarize, our contributions are three-fold:

- We propose a hierarchical and intention-aware framework to model the underlying human cognitive process to help accurately predict situated behavior in the environment.
- We design a dynamic GCN with an adaptive weight matrix that effectively captures context-aware relationships between humans and the environment.
- We conduct an empirical validation of the framework's practical applicability through an experiment in a real-world VR environment, demonstrating its robustness and effectiveness under realistic, noisy conditions.

2. Related Work

Gaze-body Correlation. The correlation between gaze and body movements during human-object interactions has been extensively explored in prior research [4, 7, 10–13, 18, 21, 24, 29, 36, 46, 47, 52, 54, 62], which can be broadly categorized into three main strands. The first line of works focuses on the correlation between gaze, hands, and objects [4, 7, 10–12, 24, 46]. They find that gaze typically pre-plans the trajectory for hand movements, while gaze fixations selectively focus on objects that are highly relevant to the current task. The second line of works analyzes gaze-head coordination [13, 18, 29, 47, 52, 54]. One of their main findings is the correlation among gaze position, head orientation, and head rotation velocity. Moreover, they also reveal that gaze movements slightly precede head movements. The third line of works explores the correlation between gaze and the full-body [21, 36, 52, 62]. For example, Sidenmark *et al*. [52] examine the patterns between gaze shifts and body movements and their temporal alignment. Hu *et al*. [21] analyze the correlation between gaze movements and human body movements. Inspired by the discovery that gaze plays a pivotal role in human-object interactions, offering valuable cues for predicting human actions, we leverage gaze information to enhance the accuracy of human-object prediction as well.

Gaze Prediction. Gaze, as an explicit representation of human visual attention, offers important information for mod-

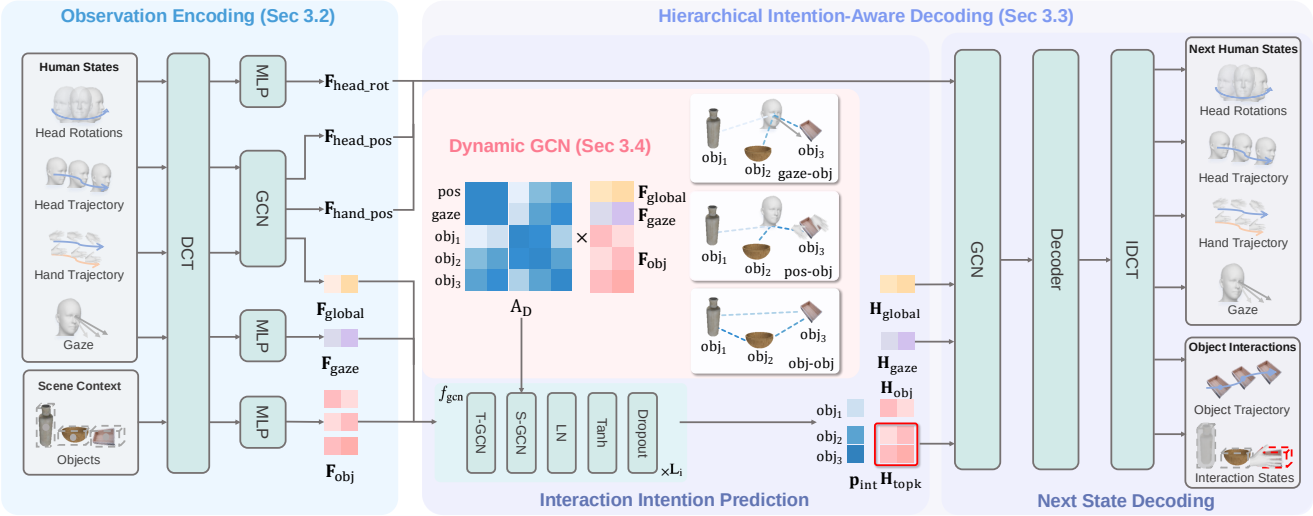


Figure 2. **Overview of our framework.** (1) an observation encoding module that captures and encodes historical human states and scene context, (2) a hierarchical intention-aware decoding module that first predicts potential top K interaction targets and then forecasts detailed human next states as well as the object interactions, and (3) a dynamic GCN that adaptively models relationships among human gaze, head positions, hand positions, and objects.

eling human intentions and behavior [16, 20, 36, 42, 60, 62]. Existing works on gaze prediction can be categorized into two groups: 2D gaze prediction [14, 18, 23, 26, 30, 31] and 3D gaze prediction [19–21, 25, 59]. 2D gaze prediction aims to estimate gaze positions in egocentric videos. Pioneer work [14] proposes the problem of predicting gaze from a 2D egocentric video. Huang *et al*. [23] tackle this problem by modeling the attention transition and predicting the saliency map for each frame. Recent work [30] leverages a transformer-based model to enhance performance. For 3D gaze prediction, the goal is to predict the gaze in a 3D scene. For example, GazeMotion [20] tries to use a CNN-based model to predict future gaze vectors and then uses the predicted gaze to help human motion prediction. Pose2Gaze [21] utilizes the historical human pose feature to predict 3D gaze. Wei et al. [59] further explored the cognitive aspects of gaze by jointly modeling attention, intentions, and tasks. We acknowledge the strong correlation between gaze and human trajectories and argue that gaze also significantly influences the next potential object to be interacted with. Therefore, we propose to simultaneously predict gaze, trajectories, and the next active object.

Next Active Object Prediction. Predicting the next active objects enables more granular user assistance [48] and has rapidly gained attention [2, 15, 17, 34, 35, 45, 50]. Furnari *et al*. [15] pioneerly propose predicting which object will be interacted with next by training a classifier. The following works [34, 35, 45] try to address this problem by predicting the interaction hotspots on 2D egocentric videos. For example, Pasca *et al*. [45] leverage pre-trained image captioning and vision-language models to extract the action context, which is then processed to forecast the next object interaction. Besides, PickAndPlace [50] introduces this problem to the robotic area and utilizes the human pose and gaze to predict the object to be picked and its place location. Ashutosh *et al*. [2] propose to anticipate all the subsequent interaction locations on the objects and corresponding human poses, given the observation of a person performing an activity. Although extensive research has been conducted on this problem, frameworks explicitly aligned with human cognition remain relatively underexplored. In contrast, we innovatively employ a cognition-aligned hierarchical framework to predict the next active objects and their future 3D trajectories, which achieves enhanced performance.

3. Method

3.1. Overview

Problem Formulation. Given historical observations of a human in an environment with objects, we aim to predict where the human will go (*trajectory*), where the human will look (*gaze*), and which objects the human will interact with (*object interaction*).

Our method utilizes historical states such as gaze, head pose, and hand/body trajectories, which are readily available on modern AR/VR devices. In virtual environments, object interactions can be directly logged from the simulation, while in physical settings, they can be inferred via depth sensing or IoT-enabled objects embedded with RFID tags [53, 63] or inertial sensors [28, 51].

Formally, we take T_h historical frames as input, consisting of two components: (1) human states, including gaze directions as unit vectors $\mathbf{g} \in \mathbb{R}^{T_\mathrm{h} \times 3}$, head orientations

$\mathbf{R}_{\text{head}} \in \mathbb{R}^{T_\text{h} \times 3 \times 3}$, head trajectory $\mathbf{P}_{\text{head}} \in \mathbb{R}^{T_\text{h} \times 3}$, and hand trajectory $\mathbf{P}_{\text{hand}} \in \mathbb{R}^{T_\text{h} \times 2 \times 3}$; (2) scene context, including object bounding boxes $\mathbf{P}_{\text{bbox}} \in \mathbb{R}^{T_\text{h} \times N \times 8 \times 3}$, object centers $\mathbf{P}_{\text{center}} \in \mathbb{R}^{T_\text{h} \times N \times 3}$, and semantic features $\mathbf{F}_{\text{clip}} \in \mathbb{R}^{N \times D_{\text{clip}}}$ derived from object labels, where N is the number of objects and D_{clip} is the semantic feature dimension. We predict the next T_f frames of human states, including gaze directions $\hat{\mathbf{g}}$, head orientations $\hat{\mathbf{R}}_{\text{head}}$, head trajectory $\hat{\mathbf{P}}_{\text{head}}$, and hand trajectory $\hat{\mathbf{P}}_{\text{hand}}$, as well as object interactions, including the trajectories of object centers $\hat{\mathbf{P}}_{\text{center}}$, and interaction states $\mathbf{p} \in \{0, 1\}^N$, indicating whether human-object interactions occur.

Architecture. As illustrated in Fig. 2, our framework consists of three core components: (1) an observation encoding module that processes historical human states and scene context into spatio-temporal features (Sec. 3.2), (2) a hierarchical intention-aware decoder that first predicts potential interaction targets and forecasts specific next human states and object interactions (Sec. 3.3), and (3) a context-aware dynamic graph convolutional network (GCN) that models human-object relationships and improves prediction accuracy (Sec. 3.4). In the following, we will detail the three modules, respectively, and introduce the training losses in Sec. 3.5.

3.2. Observation Encoding

Our observation encoding transforms historical human states ($\mathbf{g}$, $\mathbf{R}_{\text{head}}$, $\mathbf{P}_{\text{head}}$, $\mathbf{P}_{\text{hand}}$) and scene context ($\mathbf{P}_{\text{bbox}}$, $\mathbf{P}_{\text{center}}$, $\mathbf{F}_{\text{clip}}$) into spatio-temporal representations. As shown in Fig. 2, we first apply discrete cosine transform (DCT) [37, 38] to convert all input time-series data to the frequency domain following practices in human motion prediction, and then process each input through specialized encoding networks.

To encode human state, gaze vectors $\mathbf{g}$ and head rotations $\mathbf{R}_{\text{head}}$ are processed through separate multilayer perceptron (MLP) networks to produce latent embeddings $\mathbf{F}_{\text{gaze}}$ and $\mathbf{F}_{\text{head\_rot}}$, both of size $\mathbb{R}^{T_\text{h} \times D}$, where D represents the feature dimension. Head-hand trajectory $\mathbf{P}_{\text{head}}$ and $\mathbf{P}_{\text{hand}}$ are encoded using an L_e-layer spatio-temporal GCN that captures temporal dynamics and spatial relationships to get corresponding feature $\mathbf{F}_{\text{head\_pos}} \in \mathbb{R}^{T_\text{h} \times D}$ and $\mathbf{F}_{\text{hand\_pos}} \in \mathbb{R}^{T_\text{h} \times 2 \times D}$. The GCN incorporates a global node $\mathbf{F}_{\text{global}} \in \mathbb{R}^D$ to aggregate joint correspondence between hands and head across frames. Following [22], each GCN component includes a temporal GCN (T-GCN) and a spatial GCN (S-GCN). The same GCN structure is used in subsequent stages.

To encode scene context, we combine semantic priors with geometric information to represent the objects. Specifically, we process the object type labels through a pre-trained and frozen CLIP [49] encoder to extract semantic features. Object bounding box trajectory $\mathbf{P}_{\text{bbox}}$ and object center trajectory $\mathbf{P}_{\text{center}}$ are processed through MLPs to obtain spatial features. These semantic and spatial features are then fused by element-wise addition to produce object features $\mathbf{F}_{\text{obj}} \in \mathbb{R}^{T_\text{h} \times N \times D}$.

3.3. Hierarchical Intention-Aware Decoding

Drawing inspiration from cognitive processes where humans form interaction intentions before executing actions [43], we design our decoding process with a hierarchical structure. We first introduce an interaction intention prediction module (Fig. 2) to process the encoded historical human gaze patterns, trajectory cues, and scene object properties to predict the interaction probabilities of each object to select potential interaction targets. Then, we decode the next states of human and objects by integrating selected object features with updated human state features.

Specifically, the interaction intention prediction module processes three encoded streams: global pose features $\mathbf{F}_{\text{global}}$ encoding positional trajectory dynamics, gaze features $\mathbf{F}_{\text{gaze}}$ reflecting attentional focus, and object features $\mathbf{F}_{\text{obj}}$ capturing object semantics and spatial relationships. An L_i-layer GCN encoder f_{gcn} fuses the information and updates them to $\mathbf{H}_{\text{global}}$, $\mathbf{H}_{\text{gaze}}$, and $\mathbf{H}_{\text{obj}}$. The updated object features are then mapped to the interaction probabilities $\hat{\mathbf{p}}_{\text{int}} \in [0, 1]^N$ via an MLP f_p, where $\hat{}$ denotes the predicted variables:

$$\begin{aligned} \mathbf{H}_{\text{global}}, \mathbf{H}_{\text{gaze}}, \mathbf{H}_{\text{obj}} &= f_{\text{gcn}}([\mathbf{F}_{\text{global}}, \mathbf{F}_{\text{gaze}}, \mathbf{F}_{\text{obj}}]), \\ \hat{\mathbf{p}}_{\text{int}} &= \phi(f_\text{p}(\mathbf{H}_{\text{obj}})), \end{aligned} \tag{1}$$

where $[\cdot]$ denotes concatenation, and ϕ denotes the sigmoid function, mapping the GCN output to a probability distribution in range $[0, 1]$. Based on the interaction probabilities $\hat{\mathbf{p}}_{\text{int}}$, the interaction intention prediction module selects the top K object features $\mathbf{F}_\text{k} \in \mathbb{R}^{T_\text{h} \times K \times D}$ from $\mathbf{H}_{\text{obj}}$, keeping only the target candidates with the highest interaction likelihood.

The selected object features $\mathbf{F}_\text{k}$, together with the updated human state features including $\mathbf{H}_{\text{gaze}}$, $\mathbf{H}_{\text{global}}$, $\mathbf{F}_{\text{head\_pos}}$, $\mathbf{F}_{\text{hand\_pos}}$ and $\mathbf{F}_{\text{head\_rot}}$, are then passed to an L_d-layer spatio-temporal GCN decoder for next state decoding. Finally, a prediction network that integrates GCN, MLP and Inverse Discrete Cosine Transform (IDCT) decodes the detailed human-object states, including next human states with gaze $\hat{\mathbf{g}}$, head orientation $\hat{\mathbf{R}}_{\text{head}}$, head trajectory $\hat{\mathbf{P}}_{\text{head}}$ and hand trajectory $\hat{\mathbf{P}}_{\text{hand}}$, as well as object interactions including object center trajectory $\hat{\mathbf{P}}_{\text{center}}$ and refined interaction states $\hat{\mathbf{p}} \in [0, 1]^K$.

3.4. Dynamic GCN for Adaptive Human-Environment Modeling

To better model the human-environment relationships in the interaction intention prediction module, we design a dynamic

graph convolutional network (D-GCN) to model the relationship among the gaze $\mathbf{F}_{\text{gaze}}$, the joint human head and hand feature $\mathbf{F}_{\text{global}}$, and the N objects $\mathbf{F}_{\text{obj}}$, using an adaptive weight matrix $\mathbf{A}_{\text{D}} \in \mathbb{R}^{(N+2)\times(N+2)}$. Specifically, our adaptive weight matrix $\mathbf{A}_{\text{D}}$ incorporates four components: gaze-object weights $\mathbf{A}_{\text{g}} \in \mathbb{R}^{N}$, position-object weights $\mathbf{A}_{\text{p}} \in \mathbb{R}^{N}$, object-object weights $\mathbf{A}_{\text{o}} \in \mathbb{R}^{N\times N}$, and a fixed gaze-position weight of 1. See Fig. 2 for illustration.

To obtain the gaze-object weights $\mathbf{A}_{\text{g}}$, we first calculate the angle θ_{g} between the gaze direction $\mathbf{g}$ and the lines connecting the head position to the N object center position $\mathbf{d}_{\text{g}} = \mathbf{P}_{\text{center}} - \mathbf{P}_{\text{head}}$, according to Eq. (2). Then these features are fed into an MLP f_{g} to predict the weight $\mathbf{A}_{\text{g}}$ as in Eq. (3):

$$\theta_{\text{g}} = \cos^{-1}\left(\frac{\mathbf{g}\cdot\mathbf{d}_{\text{g}}}{\|\mathbf{g}\|\|\mathbf{d}_{\text{g}}\|}\right), \tag{2}$$

$$\mathbf{A}_{\text{g}} = f_{\text{g}}([\theta_{\text{g}}, \|\mathbf{d}_{\text{g}}\|]), \tag{3}$$

where $\theta_{\text{g}} \in \mathbb{R}^{T_{\text{h}}\times N}$, and $\|\mathbf{d}_{\text{g}}\| \in \mathbb{R}^{T_{\text{h}}\times N}$ is the Euclidean distance between head and N objects.

For position-object weights $\mathbf{A}_{\text{p}}$, we first calculate the minimum Euclidean distance $\mathbf{d}_{\text{p}} \in \mathbb{R}^{N}$ between each object center and the three key positions: the head $\mathbf{P}_{\text{head}}$, the left hand $\mathbf{P}^{0}_{\text{hand}}$, and the right hand $\mathbf{P}^{1}_{\text{hand}}$, averaged across T_{h} frames. Formally, the distance for i-th object $\mathbf{d}^{i}_{\text{p}}$ and the corresponding position-object weights $\mathbf{A}^{i}_{\text{p}}$ are defined as:

$$\mathbf{d}^{i}_{\text{p}} = \frac{1}{T_{\text{h}}} min\Big(\sum_{t=1}^{T_{\text{h}}} \|\mathbf{P}^{i,t}_{\text{center}} - \mathbf{P}^{t}_{\text{head}}\|, \sum_{t=1}^{T_{\text{h}}} \|\mathbf{P}^{i,t}_{\text{center}} - \mathbf{P}^{0,t}_{\text{hand}}\|, \sum_{t=1}^{T_{\text{h}}} \|\mathbf{P}^{i,t}_{\text{center}} - \mathbf{P}^{1,t}_{\text{hand}}\|\Big), \tag{4}$$

$$\mathbf{A}^{i}_{\text{p}} = \exp(-\mathbf{d}^{i}_{\text{p}}), \tag{5}$$

where $\mathbf{P}^{t}_{\text{head}}$ represents the head position at frame t, and $\mathbf{P}^{0,t}_{\text{hand}}$ and $\mathbf{P}^{1,t}_{\text{hand}}$ represent the positions of the left and right hands at frame t.

For object-object weights $\mathbf{A}_{\text{o}}$, we first compute the average Euclidean distance $\mathbf{d}_{\text{o}} \in \mathbb{R}^{N\times N}$ between each two objects across T_{h} time frames and obtain the corresponding weight $\mathbf{A}_{\text{o}}$ as in Eq. (7):

$$\mathbf{d}^{i,j}_{\text{o}} = \frac{1}{T_{\text{h}}} \sum_{t=1}^{T_{\text{h}}} \|\mathbf{P}^{i,t}_{\text{center}} - \mathbf{P}^{j,t}_{\text{center}}\|, \tag{6}$$

$$\mathbf{A}^{i,j}_{\text{o}} = \exp(-\mathbf{d}^{i,j}_{\text{o}}), \tag{7}$$

where $\mathbf{P}^{i,t}_{\text{center}}$ and $\mathbf{P}^{j,t}_{\text{center}}$ represent the center positions of the i-th and j-th objects at frame t.

In implementation, we repeat the head position N times to accomplish this operation but omit the details for clarity.

These weight components are integrated into the dynamic weight matrix $\mathbf{A}_{\text{D}}$ to enhance the node feature update process in the spatial GCN (S-GCN):

$$\mathbf{X}^{(l+1)} = \mathbf{D}^{-1/2}\mathbf{A}_{\text{D}}\mathbf{D}^{-1/2}\mathbf{X}^{(l)}\mathbf{W}^{(l)}, \tag{8}$$

where $\mathbf{X}^{(l)} = [\mathbf{F}^{(l)}_{\text{global}}, \mathbf{F}^{(l)}_{\text{gaze}}, \mathbf{F}^{(l)}_{\text{obj}}]$ are the node features at GCN layer l, $\mathbf{D}$ is the degree matrix of $\mathbf{A}_{\text{D}}$ and $\mathbf{W}^{(l)}$ are learnable parameters for the l-th layer.

3.5. Training Losses

Our training objectives directly supervise all predicted outputs. Each component of the loss corresponds to a specific prediction head in our framework:

$$\begin{aligned}\mathcal{L} = {} & \lambda_{\text{gaze}}\mathcal{L}_{\text{gaze}} + \lambda_{\text{rot}}\mathcal{L}_{\text{rot}} + \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{center}}\mathcal{L}_{\text{center}} \\ & + \lambda_{\text{int}}\mathcal{L}_{\text{int}} + \lambda_{\text{state}}\mathcal{L}_{\text{state}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}},\end{aligned} \tag{9}$$

where $\mathcal{L}_{\text{gaze}}$ measures the error between predicted and groudtruth (GT) gaze directions using cosine similarity; $\mathcal{L}_{\text{rot}}$ is L2 loss of head orientation; $\mathcal{L}_{\text{pos}}$ is L2 losses of head, hands trajectory positions, and $\mathcal{L}_{\text{center}}$ is L2 loss of object center trajectories; $\mathcal{L}_{\text{int}}$ and $\mathcal{L}_{\text{state}}$ use binary cross-entropy loss between predicted interaction probabilities and GT labels in the candidate selection and decoding stages, respectively. $\mathcal{L}_{\text{vel}}$ promotes smooth human and object trajectories by calculating L2 distance between predicted and GT velocities; The hyperparameters λ_{*} balance the contributions of these loss terms. Please refer to the supplementary material for definitions of these terms.

4. Experiment

4.1. Setup

Dataset. We train our method and conduct experiments on the Aria Digital Twin (ADT) dataset [44], captured by human participants wearing Aria glasses while performing real-world activities in two realistic indoor scenes. The dataset includes 236 sequences, 2 real indoor scenes, and 398 object geometry models. The sequences, recorded at 30 Hz, contain multi-modal data such as egocentric RGB videos, 6DoF head device poses, 3D eye gaze data, *etc*. We use 70 sequences with hand positions, splitting them randomly into 60 for training and 10 for testing. Additionally, we curate a specialized subset called "ADT-Hard", consisting of clips where object interaction states change between input and prediction periods, testing models' ability to anticipate new interactions beyond ongoing ones. We conduct all experiments on this subset as well.

For our real-world practical evaluation, we collect 6 sequences with an average duration of 1 minute from human participants who are instructed to perform natural interaction sequences in a VR environment (see Fig. 3 (d)). We apply

Method	ADT						ADT-Hard	
	Hand ↓	Head (Dist.) ↓	Head (Dir.) ↓	Gaze ↓	Object Center ↓	Object AP ↑	Object Center ↓	Object AP ↑
Pose2Gaze [21]	-	-	-	16.55	-	-	-	-
SIF3D [36]	102.70	63.92	-	-	-	-	-	-
HOIMotion [22]	79.58	45.51	-	-	-	-	-	-
GazeMotion [20]	169.30	133.10	-	14.82	-	-	-	-
PickAndPlace [50]	100.82	61.98	-	-	113.80	68.72	85.90	35.63
Ours	**78.77**	**44.15**	**8.10**	**14.02**	**88.02**	**98.56**	**84.78**	**70.49**

Table 1. **Comparison to the state-of-the-arts on ADT [44] dataset and ADT-Hard subset.** Missing entries (-) indicate that the methods do not predict the corresponding properties.

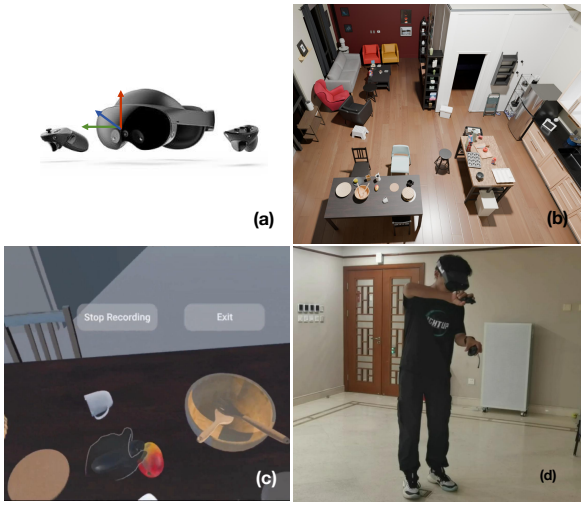


Figure 3. **Experimental setup for the real-world VR evaluation.** (a) Meta Quest Pro headset and Touch controllers used for interaction and data collection. (b) A top-down view of the interactive virtual environment. (c) The participant's egocentric view of the interactive VR scene. (d) Third-person view of a participant in the real world performing a natural interaction sequence with the virtual environment.

no preprocessing to the data to preserve authentic characteristics like sensor noise and behavioral variations, enabling a realistic test of our model's robustness.

Our method is trained on publicly available datasets and does not involve the use of facial or other privacy-sensitive biometric information. the data we used is obtained with the consent of the participants.

Real-world VR Experimental Setup. To evaluate the practical applicability of our approach, we conduct user studies in a VR environment with human participants. We use a Meta Quest Pro headset and Meta Touch controllers for the system, as shown in Fig. 3(a). The headset displays the VR interface and captures gaze vectors and 6DoF head pose, while the controllers track hand trajectories and enable interaction with virtual objects. The real-time positions and states of all virtual objects are recorded within the Unity scene during interactions.

We construct a virtual environment using object assets from the ADT dataset within the Unity Engine (v6.1), creating an interactive room scenario that mirrors real-world domestic environments, as shown in Fig. 3(b) and (c). The scene is rendered and displayed to users in a Meta Quest Pro headset.

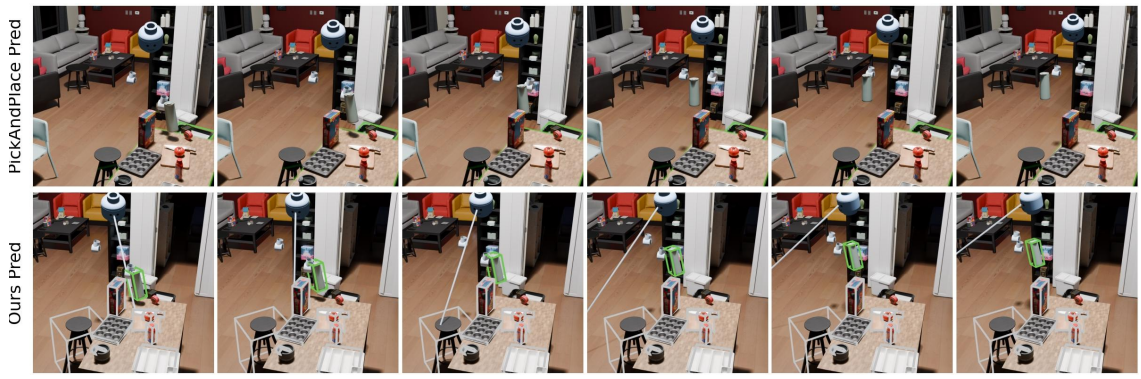
Baselines. Since no directly comparable baseline exists, we evaluate our method against five most recent related State-of-the-Art (SOTA) approaches, including Pose2Gaze [21], SIF3D [36], HOIMotion [22], GazeMotion [20], and PickAndPlace [50]. We use their official code for training and evaluating on the ADT dataset. Please refer to supplementary material for more implementation details.

Metrics. Follow common practice [3, 20–22, 36, 50], we evaluate the approaches using the standard metrics: (1) For trajectory forecasting, we report the mean L2 distance error (in millimeters) between predicted and GT positions for Hand (mm), Head (mm), and Object Center (mm); (2) For orientation prediction, we measure the average angular error in degrees for Head (°) and Gaze (°); (3) For object interaction prediction, we compute Average Precision (AP) by calculating the area under the Precision-Recall curve, varying the interaction threshold and aggregating precision-recall pairs.

Implementation Details. Our model takes historical $T_{\text{h}} = 15$ frames as input and predicts the future $T_{\text{f}} = 15$ frames, by sampling 30 Hz data with stride 2 to process 1 second of history for 1 second of future prediction. We encode $N = 48$ objects, with our hierarchical decoder selecting the top $K = 12$ objects. Our model has $\text{L}_{\text{e}} = 4$ encoding GCN layers, $\text{L}_{\text{i}} = 4$ D-GCN layers, and $\text{L}_{\text{d}} = 16$ decoding layers with $D = 16$ feature dimension. Our model achieves real-time inference at 33 FPS on a single NVIDIA 2080Ti GPU. We provide more details in Sec. B.2.

(a) Historical input observations.

(b) Ground Truth future sequences.



(c) Prediction from our model (top) and PickAndPlace (bottom).

Figure 4. **Comparison of our method with PickAndPlace [50] in a real-world VR environment.** The columns in each subfigure represent consecutive timesteps. Input sequences (a) and ground truth (b) are shown across three forms: rendered visualization, exocentric view, and egocentric view. (c) presents visualization of predictions from PickAndPlace (top) and our method (bottom). Visual elements include gaze direction (rays), head pose, hand positions (Lego figures), and interaction targets (bounding boxes). Bold bounding boxes indicate the top-K interaction candidates from our hierarchical framework, while the white-to-green gradient represents increasing predicted interaction probabilities.

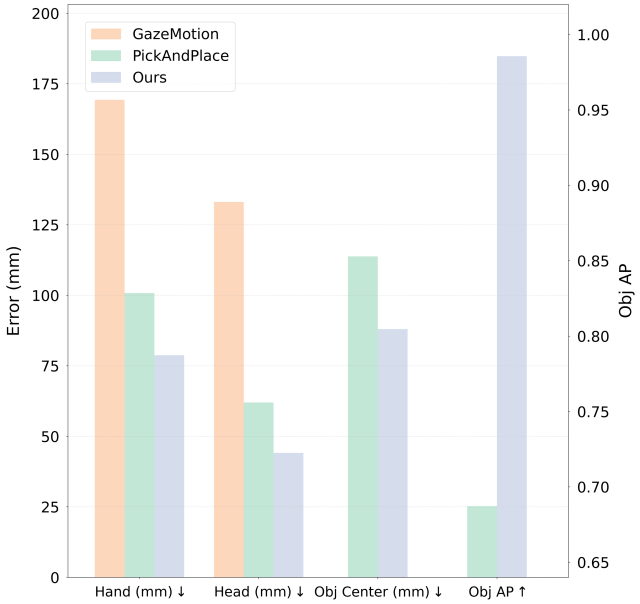


Figure 5. **Quantitative comparison of our method with PickAndPlace [50] and GazeMotion [20] on real-world VR evaluation.** We show errors on hand, head, and object center prediction as well as object interaction AP.

4.2. Real-world Experiments in a VR Environment

Quantitative Results. Fig. 5 presents the quantitative performance of our method and the most related baseline PickAndPlace [50] and GazeMotion [20] on real-world VR interaction sequences. Driven by our cognition-inspired design that hierarchically models human-object interaction intentions, our approach exhibits robust performance under noisy realistic conditions and significantly outperforms PickAndPlace [50] and GazeMotion [20] across the metrics. Notably, our method achieves a substantially higher Object Average Precision, which highlights its strength in predicting users' upcoming interaction targets. This capability is indispensable for proactive VR applications, such as intention-aware interaction assistance.

Qualitative Results. A qualitative result from a captured live interaction sequence is presented in Fig. 4, which demonstrates our method's superior predictive capabilities compared to PickAndPlace [50]. Our model successfully anticipates user gaze patterns, future pick-up behaviors, and upcoming interaction objects (a kettle), while PickAndPlace produces incorrect predictions.

Notably, This visualization provides clear evidence for the effectiveness of our cognition-aligned hierarchical framework. As demonstrated by the bold bounding boxes, the framework first generates contextually coherent candidates that are semantically and spatially relevant, providing meaningful contextual information for subsequent detailed prediction. Furthermore, as revealed by the bounding box color gradients, our final interaction probability predictions accurately reflect true user intentions. The intended target is highlighted in dark green, corresponding to a high interaction probability, while other plausible candidates remain white due to their low assigned probabilities. These results demonstrate that our method can produce accurate and reasonable predictions that can potentially support improved downstream applications such as personal assistants and adaptive game interactions. For more visualizations, please refer to our video demonstration.

4.3. Experiments on ADT dataset

While our user study confirmed the framework's real-world feasibility and provided qualitative insights into its usability, a systematic and quantitative comparison is essential to validate its algorithmic superiority. Therefore, we compare our method to state-of-the-art methods on the ADT dataset [44] and its more challenging subset ADT-Hard.

As shown in Tab. 1, GazeMotion [20] and Pose2Gaze [21] do not incorporate environmental context, which limits their performance. SIF3D [36], HOIMotion [22], and PickAndPlace [50] utilize environmental information, such as point clouds and object bounding boxes, but lack modeling of the human-environment relationship and interaction intentions, lead to reduced prediction accuracy. Our model achieves superior performance across all metrics and demonstrates a significant advantage in recognizing the next active objects on the ADT-Hard dataset, validating the robustness of our proposed cognition-aligned hierarchical decoding and dynamic weight mechanism.

Fig. 6 shows a qualitative comparison result with the baselines. Notably, even when the input gaze does not fixate on the target object, our method still accurately predicts the next active object and generates contextually appropriate situated human behavior. In contrast, PickAndPlace [50] fails to identify the interaction target and generates inaccurate interaction behavior. Other baselines [20–22, 36] do not explicitly model the human-object interaction intention. As a result, their predictions are limited to extending historical trajectories, failing to accurately identify upcoming interaction targets or predict correct human states, leading to noticeable discrepancies from the GT. Please refer to supplementary material for additional qualitative comparison and video results.

4.4. Ablation Study

Ablation of Each Module. To validate the effectiveness of the key components in our framework, we compare our approach to two ablations in Tab. 2 on the ADT [44] dataset and ADT-Hard subset. In ablation (a), we remove the entire hierarchical decoding structure, including the intention prediction module and dynamic GCN and directly decoding human states and object interactions with a vanilla GCN

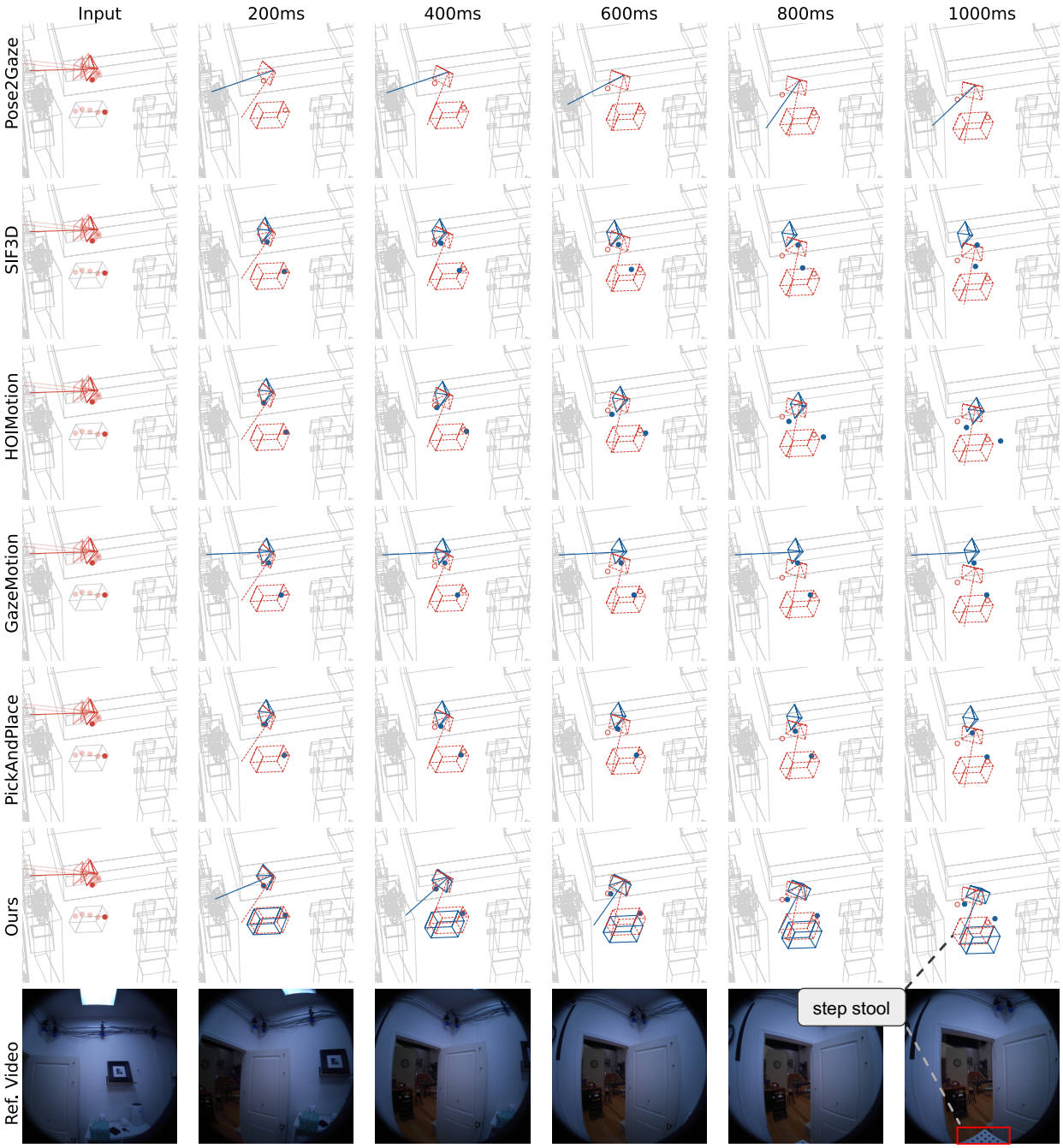


Figure 6. **Qualitative comparison of our method with state-of-the-art methods on ADT [44] dataset.** The leftmost column shows the historical input observation, where color shades from light to dark represent progression from earlier to later time, while the remaining columns display predictions at 200ms intervals from 200ms to 1000ms. In each visualization, red and blue elements represent GT and predictions, respectively. The visual elements include: gaze direction (ray), head position and orientation (pyramid), hand positions (two points), and interacted objects (bounding boxes). The interaction object type is labeled in the final frame. The bottom row shows reference frames from the first-person perspective video.

decoder of the same number of layers. In ablation (b), we replace the dynamic GCN with vanilla GCN by removing the dynamic weight matrix $\mathbf{A}_{\mathrm{D}}$.

As shown in Tab. 2, removing either module significantly

Method	ADT						ADT-Hard	
	Hand ↓	Head (Dist.) ↓	Head (Dir.) ↓	Gaze ↓	Object Center ↓	Object AP ↑	Object Center ↓	Object AP ↑
(a) *w/o* Hier. Decoding	81.16	45.10	**8.08**	14.05	97.19	97.42	97.97	60.84
(b) *w/o* Dynamic GCN	79.63	44.65	8.12	14.08	89.37	98.01	85.64	66.70
(c) **Ours**	**78.77**	**44.15**	8.10	**14.02**	**88.02**	**98.56**	**84.78**	**70.49**

Table 2. **Ablation study of our approach on ADT [44] dataset and ADT-Hard subset.** Ablation (a) removes the interaction intention prediction module (including the dynamic GCN) and directly decodes the next human states and object interactions. Ablation (b) replaces the dynamic GCN in the intention prediction module with vanilla GCN by removing the dynamic weight matrix $\mathbf{A}_D$.

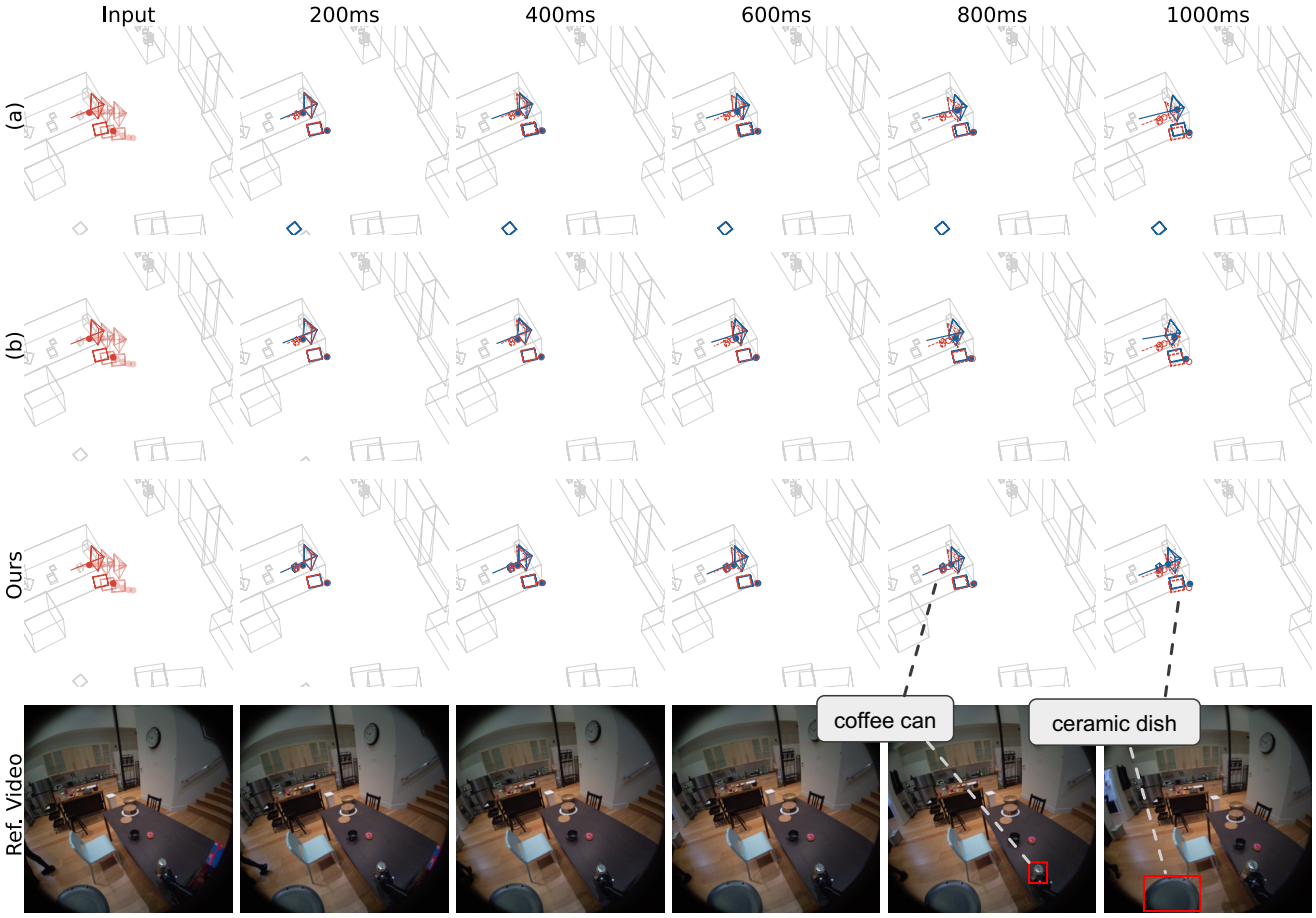


Figure 7. **Qualitative comparison of our full model with ablated baselines on ADT [44] dataset.** Ablation (a) removes the interaction intention prediction module and directly decodes the next human states and object interactions. Ablation (b) replaces the dynamic GCN with vanilla GCN by removing the dynamic weight matrix $\mathbf{A}_D$. The visualization format is kept the same as in Fig. 6.

degrades performance across most metrics. (a) The removal of hierarchical decoding leads to substantial performance drops in object prediction and AP scores, demonstrating that our approach of predicting intentions before detailed decoding is critical for accurate interaction prediction in complex scenarios. Without the (b) dynamic weight matrix, we observe increased errors in position predictions and decreased AP scores, confirming that our context-aware mechanism effectively improves human-environment relationship modeling.

Fig. 7 presents a representative qualitative comparison example from the ADT-Hard subset, illustrating the contribution of each module to accurate predictions. Our full model successfully predicts both the ceramic dish and coffee can interactions. In contrast, the model without (a) hierarchical decoding not only misses the coffee can interaction but also predicts false positives, as seen in the blue bounding box in row (a) of Fig. 7. Similarly, the model without (b) the dynamic GCN only predicts the ongoing dish interaction, but misses the new coffee can interaction. These

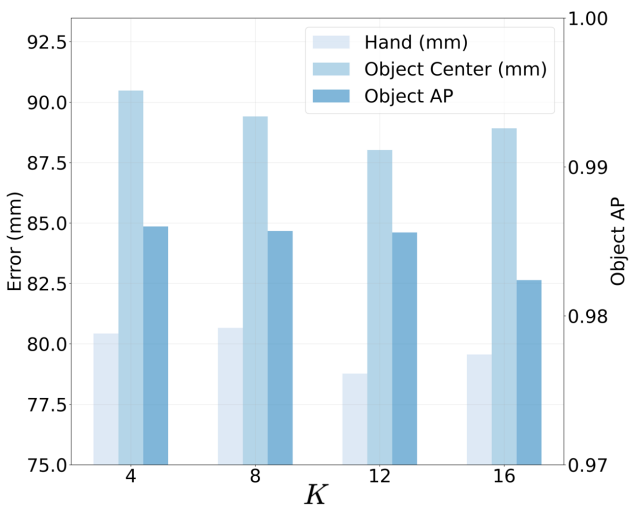


Figure 8. **Ablation study on different numbers of selected objects (K).** We show errors on hand and object center prediction as well as the object interaction AP on the ADT dataset.

differences become more pronounced in later predictions, such as at 1000ms (rightmost column). This demonstrates both components are essential for accurate interaction modeling in complex environments. Additional results are provided in the supplementary material.

Number of Selected Objects. We investigate the impact of the number of selected objects K in our hierarchical intention-aware decoding framework on the ADT dataset. Fig. 8 visualizes three key metrics for human-object interaction performance: hand position error (mm), object center error (mm), and object interaction AP. We evaluate our method with K ranging from 4 to 16 objects and find robust performance across all configurations.

As shown in Fig. 8, as K increases from 4 to 12, hand position and object center errors decrease, with a slight trade-off in object interaction AP. At $K = 16$, performance degrades, suggesting that while non-interacted objects with high interaction probability provide useful context, but excessive inclusion introduces redundancy that harms performance. We select $K = 12$ for our final model, balancing performance across metrics.

5. Conclusion

In this work, we present a cognition-inspired framework for predicting comprehensive situated human behavior, including trajectory, gaze, and object interaction. Drawing from cognitive science research, our hierarchical intention-aware decoding approach first predicts potential interaction targets before forecasting detailed human states and object interactions, mirroring how humans form intentions prior to executing actions. We develop a dynamic GCN with an adaptive weight matrix that effectively captures context-aware human-environment relationships. Extensive experiments on challenging real-world benchmark demonstrate that our approach significantly outperforms state-of-the-art methods, particularly in complex interaction scenarios where accurate intention prediction is crucial, validating our cognitively-aligned approach to situated human behavior prediction.

Limitations. While the model occasionally makes incorrect predictions, they are often reasonable due to the diversity of human behaviors. This limitation stems from the current approach's focus on a single deterministic prediction. Future work could explore modeling behavioral diversity to improve accuracy and robustness.

References

[1] Dolores Albarracín and Wenhao Dai. Chapter three - the impact of the environment on behavior. In *Advances in Experimental Social Psychology*, pages 151–201. Academic Press, 2024. 1

[2] Kumar Ashutosh, Georgios Pavlakos, and Kristen Grauman. Fiction: 4d future interaction prediction from video. *arXiv preprint arXiv:2412.00932*, 2024. 3

[3] Wentao Bao, Lele Chen, Libing Zeng, Zhong Li, Yi Xu, Junsong Yuan, and Yu Kong. Uncertainty-aware state space transformer for egocentric 3d hand trajectory forecasting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13702–13711, 2023. 6

[4] Anil Ufuk Batmaz, Aunnoy K Mutasim, Morteza Malekmakan, Elham Sadr, and Wolfgang Stuerzlinger. Touch the wall: Comparison of virtual and augmented reality with conventional 2d screen eye-hand coordination training systems. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 184–193. IEEE, 2020. 2

[5] Anna Belardinelli. Gaze-based intention estimation: principles, methodologies, and applications in hri. *ACM Transactions on Human-Robot Interaction*, 13(3):1–30, 2024. 1

[6] Valerio Belcamino, Miwa Takase, Mariya Kilina, Alessandro Carfì, Fulvio Mastrogiovanni, Akira Shimada, and Sota Shimizu. Gaze-based intention recognition for human-robot collaboration. In *Proceedings of the 2024 International Conference on Advanced Visual Interfaces*, pages 1–5, 2024. 1

[7] Jennifer K Bertrand and Craig S Chapman. Dynamics of eye-hand coordination are flexibly preserved in eye-cursor coordination during an online, digital, object interaction task. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–13, 2023. 2

[8] Myles Brand. *Intending and acting: Toward a naturalized action theory*. MIT press Cambridge, MA, 1984. 1

[9] Michael Bratman. Intention, plans, and practical reason. 1987. 1

[10] Adrien Coudiere and Frederic R Danion. Eye-hand coordination all the way: from discrete to continuous hand movements. *Journal of Neurophysiology*, 131(4):652–667, 2024. 2

[11] Anouk J De Brouwer and Miriam Spering. Eye-hand coordination during online reach corrections is task dependent. *Journal of Neurophysiology*, 127(4):885–895, 2022.

[12] S De Vries, R Huys, and PG Zanone. Keeping your eye on the target: eye–hand coordination in a repetitive fitts' task. *Experimental Brain Research*, 236(12):3181–3190, 2018. 2

[13] Yu Fang, Ryoichi Nakashima, Kazumichi Matsumiya, Ichiro Kuriki, and Satoshi Shioiri. Eye-head coordination for visual cognitive processing. *PloS one*, 10(3):e0121035, 2015. 2

[14] Alireza Fathi, Yin Li, and James M Rehg. Learning to recognize daily actions using gaze. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 314–327. Springer, 2012. 3

[15] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 49:401–411, 2017. 3

[16] Maxence Grand, Damien Pellier, and Francis Jambon. Gaipat-dataset on human gaze and actions for intent prediction in assembly tasks. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction*, pages 1015–1019, 2025. 3

[17] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 3

[18] Zhiming Hu, Congyi Zhang, Sheng Li, Guoping Wang, and Dinesh Manocha. Sgaze: A data-driven eye-head coordination model for realtime gaze prediction. *IEEE transactions on visualization and computer graphics*, 25(5):2002–2010, 2019. 2, 3

[19] Zhengxi Hu, Dingye Yang, Shilei Cheng, Lei Zhou, Shichao Wu, and Jingtai Liu. We know where they are looking at from the rgb-d camera: Gaze following in 3d. *IEEE Transactions on Instrumentation and Measurement*, 71:1–14, 2022. 3

[20] Zhiming Hu, Syn Schmitt, Daniel Haeufle, and Andreas Bulling. Gazemotion: Gaze-guided human motion forecasting. In *Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2024. 3, 6, 8, 1

[21] Zhiming Hu, Jiahui Xu, Syn Schmitt, and Andreas Bulling. Eye-body coordination during daily activities for gaze prediction from full-body poses. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2, 3, 6, 8, 1

[22] Zhiming Hu, Zheming Yin, Daniel Haeufle, Syn Schmitt, and Andreas Bulling. Hoimotion: Forecasting human motion during human-object interactions using egocentric 3d object bounding boxes. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 4, 6, 8, 1

[23] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 754–769, 2018. 3

[24] Roland S Johansson, Göran Westling, Anders Bäckström, and J Randall Flanagan. Eye–hand coordination in object manipulation. *Journal of neuroscience*, 21(17):6917–6932, 2001. 1, 2

[25] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6912–6921, 2019. 3

[26] Heecheol Kim, Yoshiyuki Ohmura, and Yasuo Kuniyoshi. Memory-based gaze prediction in deep imitation learning for robot manipulation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2427–2433. IEEE, 2022. 3

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 4

[28] Manon Kok, Jeroen D Hol, and Thomas B Schön. Using inertial sensors for position and orientation estimation. *arXiv preprint arXiv:1704.06053*, 2017. 3

[29] Rakshit Kothari, Zhizhuo Yang, Christopher Kanan, Reynold Bailey, Jeff B Pelz, and Gabriel J Diaz. Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *Scientific reports*, 10(1):2539, 2020. 2

[30] Bolin Lai, Miao Liu, Fiona Ryan, and James M Rehg. In the eye of transformer: Global–local correlation for egocentric gaze estimation and beyond. *International Journal of Computer Vision*, 132(3):854–871, 2024. 3

[31] Bolin Lai, Fiona Ryan, Wenqi Jia, Miao Liu, and James M Rehg. Listen to look into the future: Audio-visual egocentric gaze anticipation. In *European Conference on Computer Vision*, pages 192–210. Springer, 2024. 3

[32] Michael F Land. Eye movements and the control of actions in everyday life. *Progress in retinal and eye research*, 25(3): 296–324, 2006. 1

[33] Chenyi Li, Guande Wu, Gromit Yeuk-Yin Chan, Dishita Gdi Turakhia, Sonia Castelo Quispe, Dong Li, Leslie Welch, Claudio Silva, and Jing Qian. Satori: Towards proactive ar assistant with belief-desire-intention user modeling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–24, 2025. 1

[34] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 704–721. Springer, 2020. 3

[35] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3282–3292, 2022. 3

[36] Zhenyu Lou, Qiongjie Cui, Haofan Wang, Xu Tang, and Hong Zhou. Multimodal sense-informed forecasting of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2144–2154, 2024. 2, 3, 6, 8, 1

[37] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively generating better initial guesses

towards next stages for high-quality human motion prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6437–6446, 2022. 4

[38] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 474–489. Springer, 2020. 4

[39] Esteve Valls Mascaro, Daniel Sliwowski, and Dongheui Lee. HOI4ABOT: Human-object interaction anticipation for human intention reading assistive roBOTs. In *7th Annual Conference on Robot Learning*, 2023. 1

[40] Alfred R Mele. *Springs of action: Understanding intentional behavior*. Oxford University Press, 1992. 1

[41] Mahsa Nasri. Towards intelligent vr training: A physiological adaptation framework for cognitive load and stress detection. *arXiv preprint arXiv:2504.06461*, 2025. 1

[42] Süleyman Özdel, Yao Rong, Berat Mert Albaba, Yen-Ling Kuo, Xi Wang, and Enkelejda Kasneci. Gaze-guided graph neural network for action anticipation conditioned on intention. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications*, pages 1–9, 2024. 3

[43] Elisabeth Pacherie. The phenomenology of action: A conceptual framework. *Cognition*, 107(1):179–217, 2008. 1, 4

[44] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20133–20143, 2023. 5, 6, 8, 9, 10, 1, 2, 3, 4

[45] Razvan-George Pasca, Alexey Gavryushin, Muhammad Hamza, Yen-Ling Kuo, Kaichun Mo, Luc Van Gool, Otmar Hilliges, and Xi Wang. Summarize the past to predict the future: Natural language descriptions of context boost multimodal object interaction anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18286–18296, 2024. 3

[46] Jeff B Pelz and Roxanne Canosa. Oculomotor behavior and perceptual strategies in complex tasks. *Vision research*, 41 (25-26):3587–3596, 2001. 1, 2

[47] Kevin Pfeil, Eugene M Taranta, Arun Kulshreshth, Pamela Wisniewski, and Joseph J LaViola Jr. A comparison of eye-head coordination between virtual and physical realities. In *Proceedings of the 15th ACM Symposium on Applied Perception*, pages 1–7, 2018. 2

[48] Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. An outlook into the future of egocentric vision. *International Journal of Computer Vision*, 132(11):4880–4936, 2024. 3

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 4

[50] Haziq Razali and Yiannis Demiris. Using eye gaze to forecast human pose in everyday pick and place actions. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8497–8503. IEEE, 2022. 3, 6, 7, 8, 1, 2

[51] Mohammad Shahbazi, Seyed Hojat Mirtajadini, and Hamidreza Fahimi. Visual–inertial object tracking: Incorporating camera pose into motion models. *Expert Systems with Applications*, 229:120483, 2023. 3

[52] Ludwig Sidenmark and Hans Gellersen. Eye, head and torso coordination during gaze shifts in virtual reality. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(1): 1–40, 2019. 2

[53] Jongchul Song, Carl T Haas, and Carlos H Caldas. A proximity-based method for locating rfid tagged objects. *Advanced Engineering Informatics*, 21(4):367–376, 2007. 3

[54] Alexander Stamenkovic, Paul J Stapley, Rebecca Robins, and Mark A Hollands. Do postural constraints affect eye, head, and arm coordination? *Journal of neurophysiology*, 120(4): 2066–2082, 2018. 2

[55] Tong Tong, Rossitza Setchi, and Yulia Hicks. Context change and triggers for human intention recognition. *Procedia Computer Science*, 207:3826–3835, 2022. 1

[56] Tong Tong, Rossitza Setchi, and Yulia Hicks. Human intention recognition using context relationships in complex scenes. *Expert Systems with Applications*, 266:126147, 2025.

[57] Stefano Triberti and Giuseppe Riva. Being present in action: a theoretical model about the "interlocking" between intentions and environmental affordances. *Frontiers in Psychology*, 6: 2052, 2016. 1

[58] Portia Wang, Eugy Han, Anna Queiroz, Cyan DeVeaux, and Jeremy N Bailenson. Predicting and understanding turn-taking behavior in open-ended group activities in virtual reality. *arXiv preprint arXiv:2407.02896*, 2024. 1

[59] Ping Wei, Yang Liu, Tianmin Shu, Nanning Zheng, and Song-Chun Zhu. Where and why are they looking? jointly inferring human attention and intentions in complex tasks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6801–6809, 2018. 3

[60] Haodong Yan, Zhiming Hu, Syn Schmitt, and Andreas Bulling. Gazemodiff: Gaze-guided diffusion model for stochastic human motion prediction. *arXiv preprint arXiv:2312.12090*, 2023. 3

[61] Yiwen Zhang, Jianing Hao, Zhan Wang, Hongling Sheng, and Wei Zeng. Facilitating video story interaction with multi-agent collaborative system. *arXiv preprint arXiv:2505.03807*, 2025. 1

[62] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C Karen Liu, and Leonidas J Guibas. Gimo: Gaze-informed human motion prediction in context. In *European Conference on Computer Vision*, pages 676–694. Springer, 2022. 2, 3

[63] Junyi Zhou and Jing Shi. Rfid localization algorithms and applications—a review. *Journal of intelligent manufacturing*, 20:695–707, 2009. 3

We provide additional method descriptions and experimental setup in this supplementary material.

A. Method Details

A.1. Detailed Training Losses

Our training objectives consist of multiple components:

$$\begin{aligned}\mathcal{L} =&\lambda_{\text{gaze}}\mathcal{L}_{\text{gaze}} + \lambda_{\text{rot}}\mathcal{L}_{\text{rot}} + \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{center}}\mathcal{L}_{\text{center}} \\ &+\lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{int}}\mathcal{L}_{\text{int}} + \lambda_{\text{state}}\mathcal{L}_{\text{state}},\end{aligned} \tag{10}$$

$\mathcal{L}_{\text{gaze}}$ is the gaze direction loss using cosine similarity to align predicted gaze direction $\hat{\mathbf{g}}$ with ground truth $\tilde{\mathbf{g}}$, where $\tilde{}$ denotes the ground truth:

$$\mathcal{L}_{\text{gaze}} = 1 - \frac{\hat{\mathbf{g}} \cdot \tilde{\mathbf{g}}}{\|\hat{\mathbf{g}}\|\|\tilde{\mathbf{g}}\|}. \tag{11}$$

$\mathcal{L}_{\text{rot}}$ measures head rotation accuracy using mean squared error between predicted and ground truth head rotations:

$$\mathcal{L}_{\text{rot}} = \|\hat{\mathbf{R}}_{\text{head}} - \tilde{\mathbf{R}}_{\text{head}}\|^2. \tag{12}$$

$\mathcal{L}_{\text{pos}}$ computes position error by summing L2 distances between predicted and ground truth trajectories for both head and hands:

$$\mathcal{L}_{\text{pos}} = \|\hat{\mathbf{P}}_{\text{head}} - \tilde{\mathbf{P}}_{\text{head}}\|_2 + \|\hat{\mathbf{P}}_{\text{hand}} - \tilde{\mathbf{P}}_{\text{hand}}\|_2. \tag{13}$$

$\mathcal{L}_{\text{center}}$ measures object center trajectory accuracy using L2 distance between predicted and ground truth object centers:

$$\mathcal{L}_{\text{center}} = \|\hat{\mathbf{P}}_{\text{center}} - \tilde{\mathbf{P}}_{\text{center}}\|_2. \tag{14}$$

$\mathcal{L}_{\text{vel}}$ ensures trajectory smoothness by calculating L2 distances between predicted and ground truth velocities for head, hands, and object centers:

$$\begin{aligned}\mathcal{L}_{\text{vel}} =&\|\hat{\mathbf{V}}_{\text{head}} - \tilde{\mathbf{V}}_{\text{head}}\|_2 + \|\hat{\mathbf{V}}_{\text{hand}} - \tilde{\mathbf{V}}_{\text{hand}}\|_2 \\ &+\|\hat{\mathbf{V}}_{\text{center}} - \tilde{\mathbf{V}}_{\text{center}}\|_2,\end{aligned} \tag{15}$$

where $\hat{\mathbf{V}}_{\text{head}}$ and $\tilde{\mathbf{V}}_{\text{head}}$ are the predicted and ground truth head velocities, $\hat{\mathbf{V}}_{\text{hand}}$ and $\tilde{\mathbf{V}}_{\text{hand}}$ are the predicted and ground truth hand velocities, and $\hat{\mathbf{V}}_{\text{center}}$ and $\tilde{\mathbf{V}}_{\text{center}}$ are the predicted and ground truth object center velocities, respectively. For all components, velocities are computed as the difference between consecutive positions. For example, the ground truth head velocity at frame i is calculated as:

$$\tilde{\mathbf{V}}^i_{\text{head}} = \tilde{\mathbf{P}}^{i+1}_{\text{head}} - \tilde{\mathbf{P}}^i_{\text{head}}. \tag{16}$$

The velocities for hands and object centers follow the same computation pattern using their respective position variables.

$\mathcal{L}_{\text{int}}$ uses binary cross-entropy to supervise the predicted interaction probability $\hat{\mathbf{p}}_{\text{int}}$ from the Interaction Target Prediction module against the ground truth interaction states $\tilde{\mathbf{p}}$:

$$\mathcal{L}_{\text{int}} = \tilde{\mathbf{p}} \cdot \log(\hat{\mathbf{p}}_{\text{int}}) + (1 - \tilde{\mathbf{p}}) \cdot \log(1 - \hat{\mathbf{p}}_{\text{int}}). \tag{17}$$

$\mathcal{L}_{\text{state}}$ employs binary cross-entropy to align the final predicted object interaction states $\hat{\mathbf{p}}$ from the next state decoding module with the ground truth interaction states $\tilde{\mathbf{p}}$:

$$\mathcal{L}_{\text{state}} = \tilde{\mathbf{p}} \cdot \log(\hat{\mathbf{p}}) + (1 - \tilde{\mathbf{p}}) \cdot \log(1 - \hat{\mathbf{p}}). \tag{18}$$

B. Experiments

B.1. Baselines

To the best of our knowledge, no existing baselines perfectly align with our specific task. Therefore, we select the most-related works [20–22, 36, 50] and adapted them for a fair comparison. The detailed implementations are as follows.

- Pose2Gaze [21]: Pose2Gaze [21] defines the task of pose-based eye gaze prediction. Given a sequence of head directions $H \in \mathbb{R}^{3\times T}$ (3D unit vectors) and body joint positions $P \in \mathbb{R}^{3\times J\times T}$ over T steps, it can predict gaze sequence $G \in \mathbb{R}^{3\times T'}$ for the future T' frames. We replace the full-body pose input with the trajectories of the head and both hands, so J becomes 3 and P becomes size $3 \times 3 \times T$, while all other inputs and outputs remain unchanged. We also match the frame length and frame rate to ours: 15 input frames and 15 output frames at 15 fps. Keeping the model architecture unchanged, we retrain and test Pose2Gaze [21] on the ADT dataset [44].
- GazeMotion [20]: Continuing the line of Pose2Gaze [21], GazeMotion [20] defines gaze-guided human motion forecasting task. Given past T frames of 3D human pose $P \in \mathbb{R}^{T\times 3\times J}$ and corresponding eye-gaze unit vectors $G \in \mathbb{R}^{3\times T}$, GazeMotion [20] first predicts future gaze $G' \in \mathbb{R}^{T'\times 3}$, then utilizes this gaze information to predict future pose $P' \in \mathbb{R}^{T'\times 3\times J}$. Similarly, we adjust both its pose input and output to trajectories of human head and two hands (i.e. $J = 3$), and match the frame length and frame rate to ours. The model is retrained and evaluated on the ADT dataset [44].
- HOIMotion [22]: HOIMotion [22] leverages surrounding objects in the environment to aid in future human pose prediction. Specifically, given a historical sequence of body poses $P \in \mathbb{R}^{3\times J\times T}$, head orientations $H \in \mathbb{R}^{3\times T}$ (3D unit vectors), and 3D bounding boxes of scene objects (3D positions of the bounding box's eight vertices $o \in \mathbb{R}^{3\times 8\times M}$), it predicts future pose $P' \in \mathbb{R}^{3\times J\times T'}$ using a GCN-based encoder-residual-decoder architecture. We also adapt its pose representation to trajectories of human head and hands, setting J to 3, with its other input modalities remaining unchanged. We conduct retraining and evaluation on ADT dataset [44].
- SIF3D [36]: Unlike HOIMotion [22], SIF3D [36] utilizes 3D point clouds to represent information about the surrounding environment. It uses this point clouds data $S \in \mathbb{R}^{N\times 3}$, along with historical pose $P \in \mathbb{R}^{3\times J\times T}$ and gaze points sequences $G \in \mathbb{R}^{3\times T}$ within the cloud,

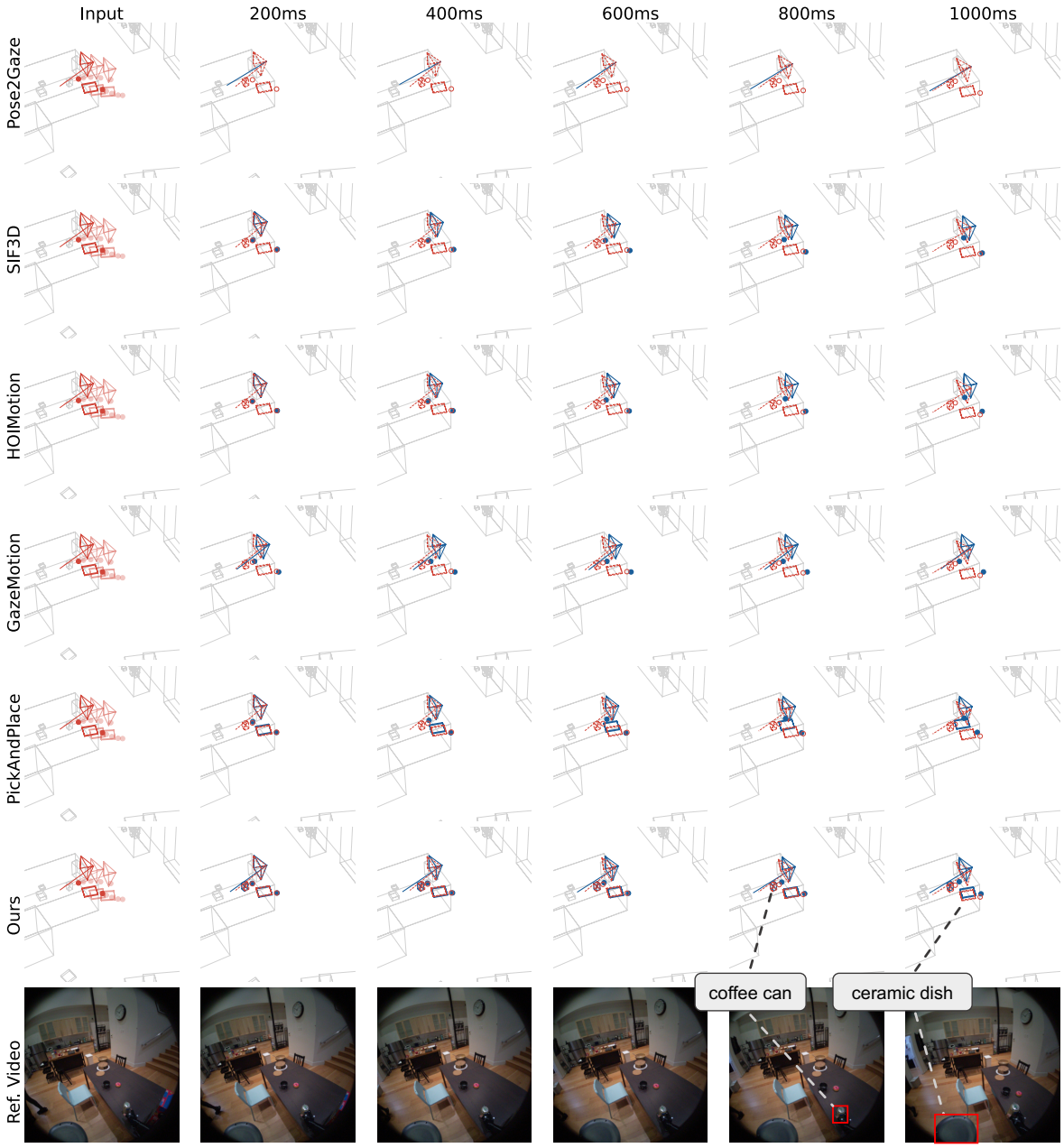


Figure 9. **Additional qualitative comparison of our method with state-of-the-art methods on ADT [44] dataset.** The leftmost column shows the historical input observation, where color shades from light to dark represent progression from earlier to later time, while the remaining columns display predictions at 200ms intervals from 200ms to 1000ms. In each visualization, red and blue elements represent GT and predictions, respectively. The visual elements include: gaze direction (ray), head position and orientation (pyramid), hand positions (two points), and interacted objects (bounding boxes). The interaction object type is labeled in the final frame. The bottom row shows reference frames from the first-person perspective video.

to predict future pose $P' \in \mathbb{R}^{3\times J\times T'}$. Similar to the aforementioned baselines, we simplify the representation of pose by using the trajectories of the person's head and hands, keep other modalities unchanged, retrain and evaluate this model on ADT dataset [44].

- PickAndPlace [50]: The goal of PickAndPlace [50] is to utilize human pose and gaze to predict an object of interest, and subsequently, the future pose. More specifically, given human pose $P \in \mathbb{R}^{3\times J\times T}$ and gaze data $G \in \mathbb{R}^{3\times T}$ over past T frames, as well as the labels and

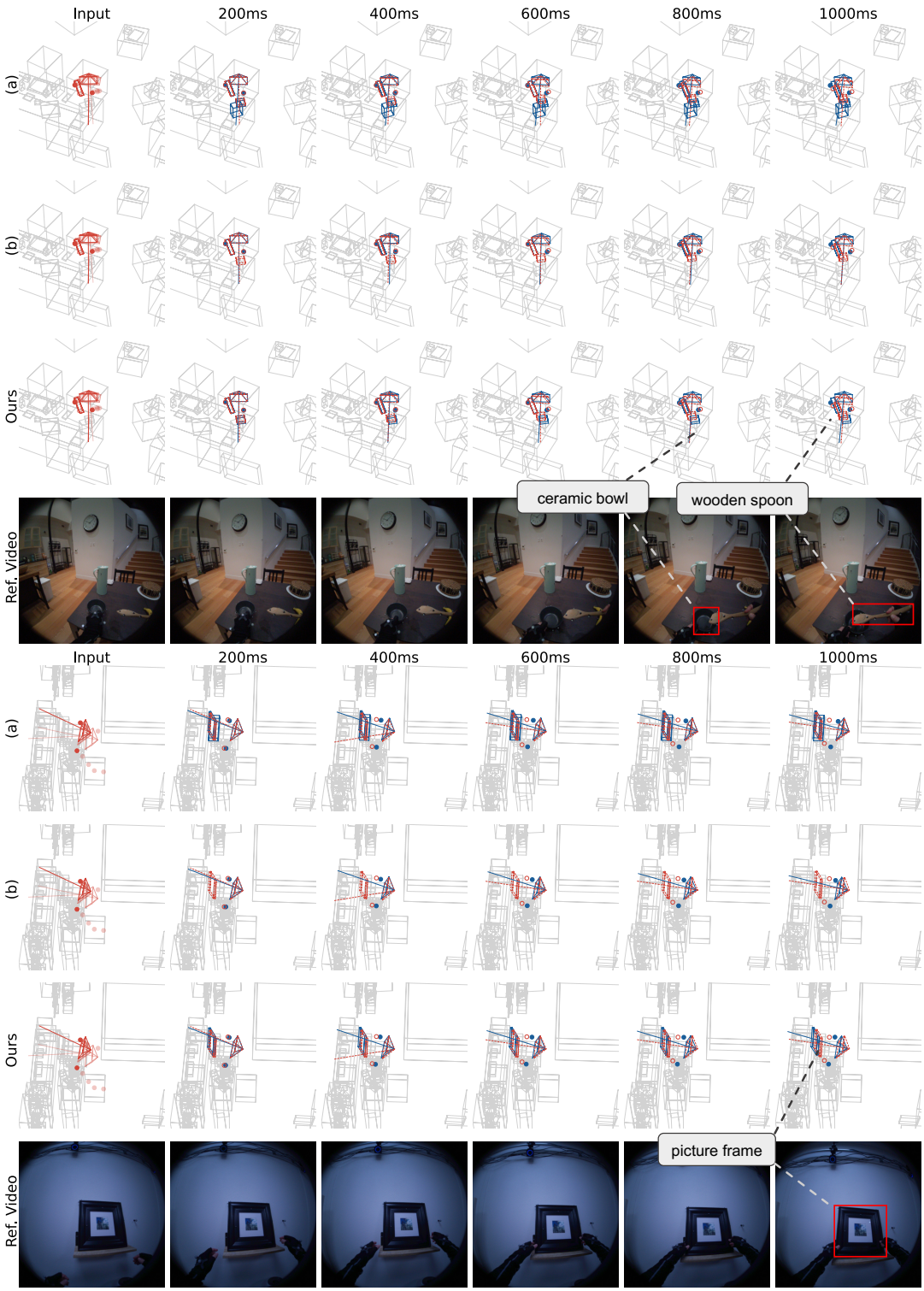


Figure 10. **Additional qualitative comparison of our full model with ablated baselines on ADT [44] dataset.** Ablation (a) replaces the dynamic GCN with vanilla GCN by removing the dynamic weight matrix $\mathbf{A}_D$. Ablation (b) removes the interaction intention prediction module and directly decodes the next human states and object interactions. The visualization format is kept the same as in Fig. 9.

coordinates of N surrounding objects, PickAndPlace [50] outputs a score for each object, representing the probabil-

ity that a human will pick it up within next T' frames. It can also predict the future human pose $P' \in \mathbb{R}^{3\times J\times T'}$. Following the aforementioned baselines, we modified the human pose data in the input and output, replacing it with the trajectories of the person's head and hands. We also redefine the meaning of the object score that this model outputs for each object, so that it represents the probability of the person interacting with the object within the next T' frames, consistent with the setting of our task. After these minor modifications, we retrain and test the model on the ADT dataset [44].

B.2. Implementation Details

Our training loss hyperparameters are set to $\lambda_{\text{gaze}} = 5.0$, $\lambda_{\text{rot}} = 5.0$, $\lambda_{\text{pos}} = 1.0$, $\lambda_{\text{center}} = 10.0$, $\lambda_{\text{vel}} = 1.0$, $\lambda_{\text{int}} = 1.0$, and $\lambda_{\text{state}} = 1.0$. We train our model using Adam optimizer [27] with an initial learning rate of 0.01 and a step learning rate scheduler with a decay factor of 0.95 per epoch. And the training is performed on 2 NVIDIA Tesla V100 GPUs for 50 epochs with a batch size of 128, taking approximately 4 hours to converge.

B.3. Additional Qualitative Results

We present additional qualitative results comparing our method with state-of-the-art baselines (Fig. 9) and ablated baselines (Fig. 10). These results clearly demonstrate the effectiveness of our cognition-based hierarchical decoding framework for situated human behavior prediction tasks. Please refer to the attached anonymous project page in the supplementary material for more video results.

In the input part of the clip represented in Fig. 9, the subject's left hand is holding a dish, and both the head and left hand trajectories are moving closer to the coffee can on the table. Our model correctly identifies not only that the dish in the left hand will maintain its interaction but also that the right hand will interact with the can. PickAndPlace, however, only determines that the plate in the left hand will continue to interact and fails to predict the interaction with the coffee pot. Furthermore, our model surpasses other baselines in predicting head and hand trajectories.

In the two cases illustrated in Fig. 10, ablation study (a), where the dynamic GCN is replaced with a vanilla GCN, resulted in omissions in the prediction of object interaction. Conversely, ablation study (b), where the interaction intention prediction module is removed, led to the prediction of incorrect objects. Our full model, in addition to predicting the correct next active object, also successfully predicts trajectories and gaze closely resemble the ground truth.