


## Article

# A Human Intention and Motion Prediction Framework for Applications in Human-Centric Digital Twins

Usman Asad <sup>1,\*</sup> , Azfar Khalid <sup>2</sup> , Waqas Akbar Lughmani <sup>3</sup> , Shummaila Rasheed <sup>1</sup>   
and Muhammad Mahabat Khan <sup>1</sup> 

<sup>1</sup> Department of Mechanical Engineering, Capital University of Science and Technology, Islamabad 45750, Pakistan; shummaila@cust.edu.pk (S.R.); drmahabat@cust.edu.pk (M.M.K.)

<sup>2</sup> Digital Innovation Research Group, Department of Engineering, School of Science & Technology, Nottingham Trent University, Nottingham NG11 8NS, UK; azfar.khalid@ntu.ac.uk

<sup>3</sup> Department of Engineering, Birmingham City University, Birmingham B4 7XG, UK; waqas.lughmani@bcu.ac.uk

\* Correspondence: usman.asad@ceme.nust.edu.pk

## Abstract

In manufacturing settings where humans and machines collaborate, understanding and predicting human intention is crucial for enabling the seamless execution of tasks. This knowledge is the basis for creating an intelligent, symbiotic, and collaborative environment. However, current foundation models often fall short in directly anticipating complex tasks and producing contextually appropriate motion. This paper proposes a modular framework that investigates strategies for structuring task knowledge and engineering context-rich prompts to guide Vision–Language Models in understanding and predicting human intention in semi-structured environments. Our evaluation, conducted across three use cases of varying complexity, reveals a critical tradeoff between prediction accuracy and latency. We demonstrate that a Rolling Context Window strategy, which uses a history of frames and the previously predicted state, achieves a strong balance of performance and efficiency. This approach significantly outperforms single-image inputs and computationally expensive in-context learning methods. Furthermore, incorporating egocentric video views yields a substantial 10.7% performance increase in complex tasks. For short-term motion forecasting, we show that the accuracy of joint position estimates is enhanced by using historical pose, gaze data, and in-context examples.

**Keywords:** human digital twin; human intention prediction; human motion generation; human–robot collaboration; artificial intelligence



Academic Editor: Junzhi Yu

Received: 13 August 2025

Revised: 11 September 2025

Accepted: 28 September 2025

Published: 1 October 2025

**Citation:** Asad, U.; Khalid, A.; Lughmani, W.A.; Rasheed, S.; Khan, M.M. A Human Intention and Motion Prediction Framework for Applications in Human-Centric Digital Twins. *Biomimetics* **2025**, *10*, 656. <https://doi.org/10.3390/biomimetics10100656>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The paradigm of digital twins has found widespread application in optimizing performance across diverse industrial sectors, including manufacturing, aerospace, and energy [1,2]. However, conventional digital twins often prioritize technical system aspects, frequently overlooking the human element crucial for Industry 5.0. Human-Centric Digital Twins (HCDTs) aim to address this by integrating models of human behavior, cognition, and even emotion, enabling personalized and adaptive feedback to human operators [3]. A truly effective HCDT allows the predicted intention and motion to be utilized in several ways: for real-time monitoring of operator performance; for providing semantic assistance such as prompting the operator on the next step or flagging procedural errors; and for enabling a robotic assistant to proactively provide physical support. Such HCDTs promise to

elevate safety, efficiency, and productivity in human–machine systems while concurrently enhancing user experience and satisfaction [4]. Traditional automation systems exhibit significant rigidity in High-Mix–Low-Volume (HMLV) industrial contexts, necessitating substantial human intervention for customization and adaptation. The advent of Machine Learning (ML), Artificial Intelligence (AI), and advanced automation technologies offers a pathway towards more flexible solutions, particularly for Small and Medium Enterprises (SMEs). These innovations can facilitate partial automation of labor-intensive tasks or enable collaborative human–robot workflows in which humans focus on complex reasoning and dexterous manipulation while robots manage repetitive actions [5,6]. Digital twins of these collaborative processes can further enhance skill transfer and enable remote oversight via Virtual Reality / Augmented Reality (VR/AR) [7,8]. Despite these technological strides, industrial robotics largely remains confined to preprogrammed caged systems, with collaborative applications still in their infancy. Recent breakthroughs in behavior cloning [9], diffusion policies [10], high-fidelity simulators [11,12], and the proliferation of Large Language Models (LLMs) [13,14] and Vision–Language Models (VLMs) [15] have expanded the horizons for semantic understanding, task planning, and perception in robotic systems [16,17].

Achieving robust spatial intelligence, such as forecasting detailed 3D human motion from visual inputs in partially structured settings, remains an important biomimetic challenge. This requires emulating the innate human ability to predict others' future actions by synthesizing a rich stream of multimodal cues. For instance, when a person glances at an empty cup and then stands up, this indicates a likely intention to get a drink. Such social cognition relies on diverse sensory inputs, including body language, posture, tracking gaze direction, recalling recent actions to understand context, and applying prior knowledge about the task and environment. In order to create machines that can seamlessly and safely work alongside people, artificial systems must be able to emulate this predictive ability. Additionally, translating this understanding into accurate 3D motion prediction remains an open problem. This predictive capability is paramount for robots to transition from reactive assistance to smart collaboration.

To address this challenge, in this paper we propose a modular and scalable framework that utilizes Vision–Language Models (VLMs) with motion diffusion techniques to predict human intent and generate plausible future human motion in semi-structured environments. Instead of pursuing end-to-end training, this approach emphasizes the integration of existing pretrained perception and reasoning modules. This design philosophy not only makes the system more practical for SMEs with limited resources but also ensures adaptability, as overall performance can improve with advancements in individual modules. To validate this framework, three use case scenarios with varying levels of complexity are utilized, including scenarios from existing datasets such as HaVID [18] and EgoExo4D [19]. This provides a systematic evaluation of the framework's ability to perform context-driven and intention-aware motion prediction. The remainder of this paper is structured as follows: Section 2 reviews related research; Section 3 details the proposed framework; Section 4 presents the validation approach and experimental results; finally, Section 6 discusses the implications of this work and outlines future directions.

## 2. Background and Related Work

The endeavor to create proactive robotic assistants capable of anticipating human needs and actions draws upon several interconnected research areas. This section reviews pertinent work in human intent prediction for Human–Robot Collaboration (HRC), AI-driven human motion generation, the role of LLMs in robotics, and the critical aspect of which datasets are available for these tasks.

### 2.1. Human Intent Prediction for Collaborative Environments

Anticipating human intent is pivotal to enhancing safety, security, ergonomics, and effective collaboration in Industry 5.0. Early work often focused on trajectory prediction in constrained scenarios [20]. More recent approaches have sought to infer higher-level goals. For instance, Huang et al. [21] proposed a hierarchical intention-tracking framework for assembly tasks, utilizing OpenPose and Kalman filtering to track wrist positions and infer both high-level task goals and low-level current actions. Similarly, Mangin et al. [22] developed hierarchical planners that infer human goals using partially observable Markov decision processes for procedural tasks. These methods underscore the importance of structured task understanding.

The integration of richer sensory data and more sophisticated AI models is an ongoing trend. Zhong et al. [23] introduced a framework for human–robot task handover that fuses a hierarchical human digital twin with deep domain adaptation while leveraging spatiotemporal graph convolutional networks. Ding et al. [24] proposed a dynamic scenario-enhanced network for predicting stochastic motions in customized assembly tasks, highlighting the need to handle variability. The use of egocentric data has also gained traction, with works like that of Mascaro et al. [25] developing intention-conditioned hierarchical architectures for long-term action anticipation based on the Ego4D dataset [26]. However, many existing methods focus on recognizing intent from past actions rather than on proactively forecasting future motion linked to that intent. They also often lack robust integration with real-time simulation for physical plausibility.

### 2.2. AI-Driven Virtual Human Motion Generation

Generating realistic and controllable human motion is a long-standing challenge in computer graphics and AI. The field has historically relied on motion capture (MoCap) datasets, which provide high-fidelity kinematic data for a wide range of human activities. A significant advancement was the creation of large-scale datasets such as HumanML3D that pair MoCap data with natural-language descriptions [27]. Pioneering works such as Adversarial Motion Priors (AMP) [28,29] trained Reinforcement Learning (RL) policies to perform tasks while using a discriminator to ensure that the resulting motions were realistic and stylistically similar to MoCap examples. This has been extended by models such as Adversarial Skill Embeddings (ASE) [30] and Conditional Adversarial Latent Models (CALM) [31], which focus on learning a low-dimensional latent space that can be sampled to direct a character's behavior.

With the advent of kinematic diffusion models, the Human Motion Diffusion Model (MDM) [32] has demonstrated a remarkable ability to generate complex motions from text prompts alone. However, because they lack physical grounding, these purely kinematic approaches often produce physically implausible motions with artifacts such as foot-sliding, floating, and ground penetration. PhysDiff [33] introduced a novel physics-guided approach that incorporates a physics simulator directly into the diffusion process, thereby correcting the motion to adhere to physical constraints.

Recent work has focused on creating unified controllers that combine the strengths of generative models with the realism of physics-based simulation for more interactive and multimodal control. MaskedMimic [34] presented a unified framework that formulates physics-based character control as a versatile motion inpainting problem. It trains a single controller to synthesize physically plausible full-body motions from partial or “masked” inputs, which can include any combination of target joint positions, text commands, or object interactions. Another state-of-the-art approach is CLoSD [35], which closes the loop between motion planning and execution. It uses a real-time autoregressive Diffusion Planner (DiP) to generate kinematic motion plans on-the-fly, which are then executed by a

robust RL-based tracking controller in a physics simulator. However, these MoCap-based methods often lack the rich scene-interaction context necessary for robustly forecasting actions in unstructured real-world environments.

### *2.3. The Role of LLMs in Intelligent Robots*

The advent of LLMs has opened up new avenues for enhancing the semantic understanding and planning capabilities of robotic systems. LLMs can parse natural language instructions, reason about task goals, and even generate plans or code snippets for robotic execution [14,16]. For instance, SayCan [17] demonstrated how LLMs can propose high-level actions grounded by pretrained robotic skills. Singh et al. [16] developed ProgPrompt using programmatic LLM prompts to generate executable plans. Ha et al. [36] explored using LLMs to guide high-level planning for data collection, then distilling this into visuomotor policies. Further, efforts such as “Asking Before Action” by Chen et al. [37] empower LLM agents to proactively seek information. These works highlight LLMs’ potential in interpreting complex instructions and decomposing them into actionable steps. However, they are primarily designed to interpret high-level commands and map them to a robot’s own action space. They are not inherently structured to analyze a continuous stream of human motion and environmental context to proactively predict a human’s future intent and generate a corresponding physically plausible motion trajectory, which is the central focus of our work. The Magentic-One framework [38] showcases the potential of multi-agent systems driven by LLMs for complex task solving. However, grounding LLM outputs in the physical world, ensuring feasibility, and integrating them with continuous motion generation remain active research areas.

### *2.4. Datasets for Human Motion and Intent*

The development of robust models for intention-based motion generation is intrinsically linked to the availability of suitable datasets. Motion capture (MoCap) datasets such as AMASS [39] and HumanML3D [27] offer precise 3D human kinematics. These datasets are invaluable for learning the fundamental dynamics and stylistic nuances of human movement. However, they are typically recorded in controlled laboratory settings, and often lack the rich object interactions and complex environmental context necessary for inferring high-level human intent. While some datasets, such as KIT Motion-Language [40], pair motion with textual descriptions, these descriptions usually label the overt action rather than the underlying intent or the broader task goal. Consequently, their utility for training models that predict intent from a wider range of cues is limited. Conversely, video-based datasets such as Ego4D [26], EgoExo4D [19], Assembly101 [41], and HaVID [18] provide a wealth of visual context, capturing humans performing tasks in more natural and often cluttered environments. Ego4D, with its extensive collection of egocentric video, offers a first-person perspective on daily activities. EgoExo4D complements this with synchronized exocentric views, providing a more holistic understanding of human actions and interactions within a scene. Assembly101 focuses specifically on procedural assembly tasks, offering structured sequences of actions. HaVID [18] contributes a dataset focused on human assembly with an emphasis on comprehensive knowledge understanding, including granular action annotations. These datasets are more conducive to understanding human–object interactions and inferring short-term goals or intentions from visual cues and task progression. However, extracting precise full-body 3D human motion comparable to MoCap quality from these “in-the-wild” videos remains a significant challenge, often relying on human pose estimation via methods such as Multi-HMR [42].

This reveals a critical gap in the lack of a unified framework capable of utilizing rich multimodal human observations such as visual data, pose dynamics, and gaze together with formal task knowledge. A bridge to connect the semantic long-term understanding derived from these inputs with a physically plausible forecast of future movements is currently lacking.

The present work directly addresses this gap by proposing a novel modular framework that creates a pipeline using the necessary context from these specific multimodal inputs for semantic intent prediction, ultimately providing context-aware motion synthesis. This approach provides a solution for proactive forecasting, a necessary capability for enabling symbiotic human–machine systems in complex environments.

### 3. Proposed Framework

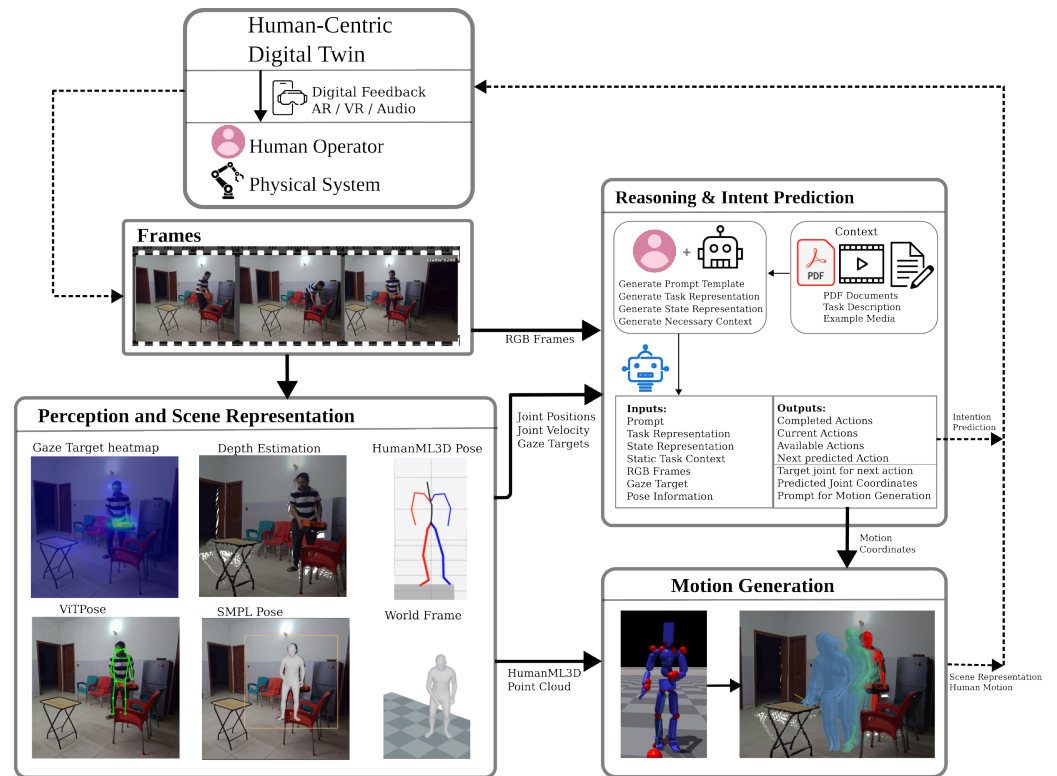
A modular framework is proposed to enable intelligent HRC by predicting human intention and future pose. This methodology is designed to bridge the gap between short-term motion forecasting and long-term action planning, ensuring that predictions are both physically plausible and contextually appropriate.

Let the input at any given time  $t$  consist of a history of video frames  $\mathcal{V} = \{I_k\}_{k=1}^t$ , extracted human poses  $\mathcal{P} = \{P_k\}_{k=1}^t$ , and static high-level task knowledge  $\mathcal{C}_{task}$ . The framework, represented by a function  $\mathcal{F}$ , processes this context to provide three outputs: the predicted task state  $S_t$ , a forecast of future hand positions  $\hat{H}_{t+\Delta t} = ((x_l, y_l), (x_r, y_r))$  for a time horizon  $\Delta t$ , and a full-body 3D motion trajectory  $\hat{M}_{t_i} = \{\hat{x}_{t_i+1}, \dots, \hat{x}_{t_i+N_{pred}}\}$ .

As shown in Figure 1, the proposed framework has three primary modules: a Perception and Scene Representation Module that processes raw visual data; an AI-driven Reasoning and Prediction Module that infers task state and forecasts future actions; and a downstream Motion Generation Module that synthesizes full-body 3D motion. The specific functions of each of these modules are detailed in Sections 3.1–3.3. This modularity allows for the individual modules to be upgraded as technology advances. While the framework provides the necessary predictive outputs for a complete Human-Centric Digital Twin (HCDDT), the implementation of feedback and physical assistance mechanisms is designated as future work.

#### 3.1. Perception and Scene Representation

The framework adopts the Skinned Multi-Person Linear Model with Extensions (SMPL-X) [43], which provides a parametric model for body shape  $\beta$ , pose  $\theta$ , and expressive hand and face parameters  $\psi$ . To obtain the pose parameters from monocular RGB video, we employ World-Grounded Human Motion Recovery via Gravity-View Coordinates (GVHMR) [42,44]. GVHMR reconstructs human pose in a global gravity-aligned coordinate system. For kinematic motion generation, the SMPL-X poses are subsequently converted into the representation utilized by HumanML3D [27]. This format captures frame-relative information such as root angular and linear velocities  $(\dot{r}_a, \dot{r}_x, \dot{r}_z)$ , root height  $(r_y)$ , and local joint positions, rotations, and velocities  $(j_p, j_r, j_v)$ , along with foot contact features  $(f)$ . This relative encoding facilitates the generation of smooth and continuous motions.



**Figure 1.** Conceptual overview of the proposed modular framework for Human Intention Prediction-based Motion Generation. Information flows from the physical system through the perception and reasoning modules to generate a final motion prediction.

Additionally, this module is required to convert the VLM's 2D predictions into 3D world coordinates. To convert 2D predictions into 3D locations, we combine dense camera-frame 3D from UniK3D [45] with the rigid transforms produced by GVHMR. Figure 2 summarizes the process. UniK3D returns per-pixel 3D points  $\mathbf{X}_c(u, v) \in \mathbb{R}^3$  for the input image. Additionally, GVHMR is used to find the transformation that maps the camera-frame to a gravity-aligned human-centric frame:  $\mathbf{X}_h = \mathbf{R}_{c \rightarrow h} \mathbf{X}_c + \mathbf{T}_{c \rightarrow h}$ . We render the camera frame SMPL mesh (from GVHMR) together with the scene point cloud (from UniK3D). An offset is estimated to co-locate the human in both modalities. We obtain the pelvis pixel  $(u_p, v_p)$  from ViTPose (used by GVHMR) and sample UniK3D at  $(u_p, v_p)$  to obtain the 3D coordinates of the pelvis position  $\mathbf{p}_c$ . We compare  $\mathbf{p}_c$  with the pelvis position of the SMPL mesh in the camera frame and apply the resulting transformation to align the mesh with the point cloud. In our experiments, minor manual refinements of this alignment were applied as needed. Given a 2D point of interest  $(u, v)$ , we compute

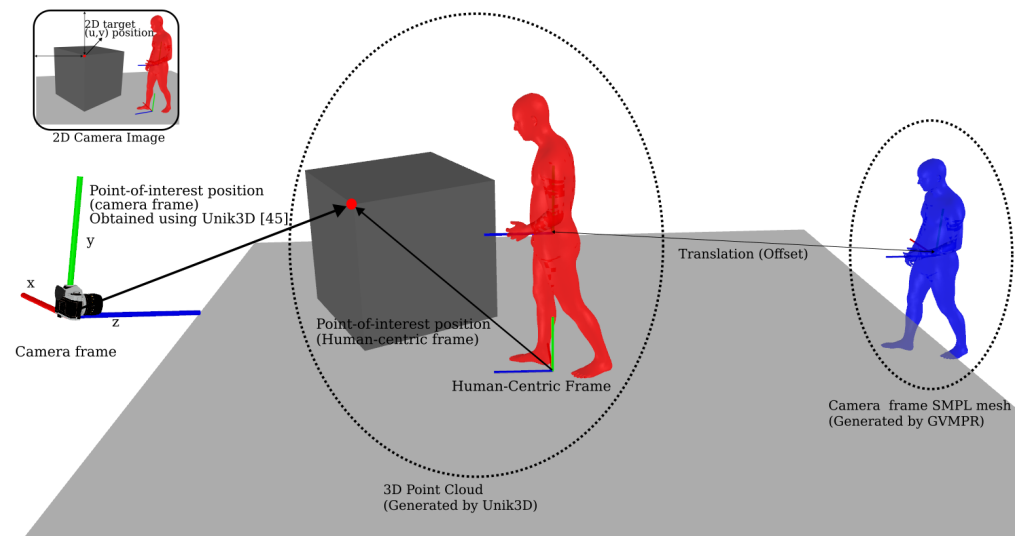
$$\mathbf{X}_h = \mathbf{R}_{c \rightarrow h} \mathbf{X}_c(u, v) + \mathbf{T}_{c \rightarrow h}. \quad (1)$$

The resulting 3D target is passed to the motion module as the goal position for a selected joint (e.g., wrist or pelvis), as further described in Section 3.3.

For specialized applications requiring different or higher-precision grounding, this module can be extended to incorporate other engineered approaches, such as processing fiducial markers.

This module also incorporates human gaze target analysis, which provides a cue for immediate intent. A change in the operator's gaze target is often a useful indicator of a change in intention, and the location of the gaze target frequently correlates with the area or object that the operator intends to interact with next. Techniques such as Gaze-LLE [46]

can be employed for gaze target tracking to further enrich the context provided to the reasoning and intent prediction module.

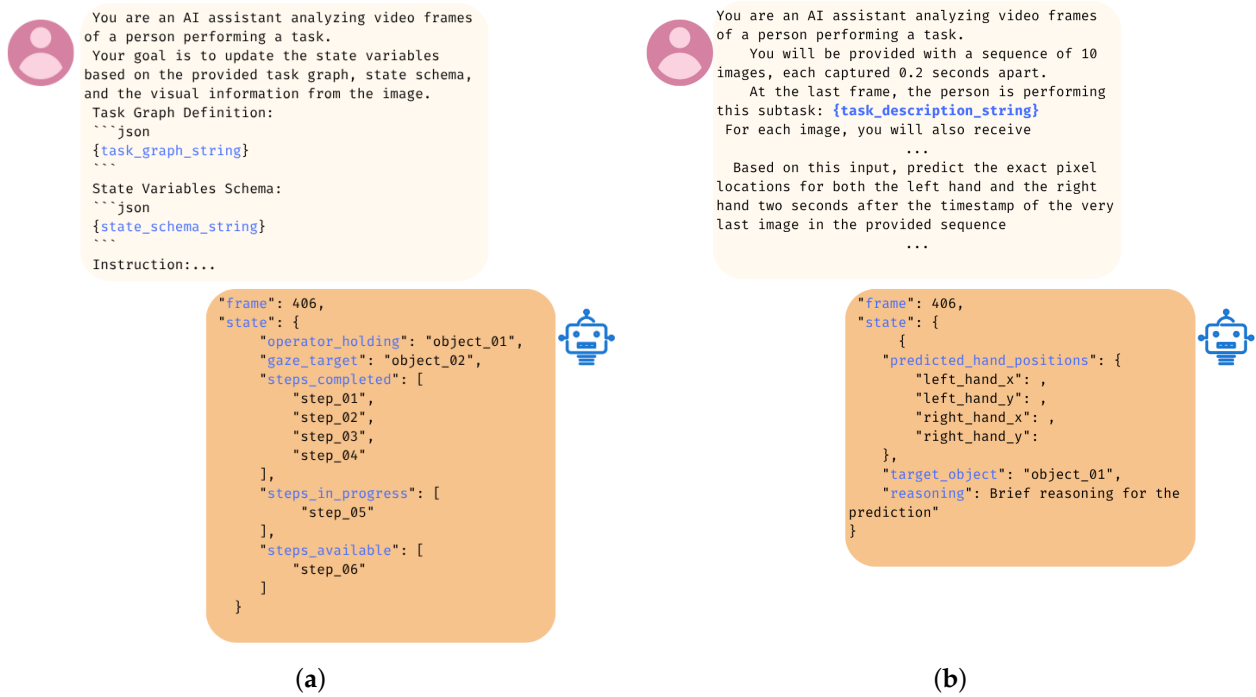


**Figure 2.** The 3D grounding process combines dense 3D from Unik3D [45] with the camera-to-world transform from GVHMR [44] to convert 2D pixel coordinates into 3D world coordinates.

### 3.2. AI-Driven Reasoning and Intent Prediction

This module serves as the cognitive core of the framework. The integration of VLMs for high-level reasoning about tasks and intent requires appropriate data structures. Task dependencies are modeled using formalisms such as Directed Acyclic Graphs (DAGs), which define the valid sequence of actions and states. Alternative formalisms, such as Planning Domain Definition Language (PDDL) [16], Knowledge Graphs (KGs), or Behavior Trees (BTs) [17,47], can also be integrated. For any use case, a common operational flow is followed. Initially, task-specific knowledge such as assembly instructions, procedural steps, or an example video is provided to the system. This involves a collaborative effort in which an AI model generates an initial draft of a task representation (i.e., an action graph and state schema), which is then refined by human experts. This task representation forms a critical part of the context, encompassing the overall goal, a mechanism for state tracking, decomposition into smaller actions, and their respective dependencies.

The prediction process follows a two-phase approach. The VLM analyzes video frames and appropriate context from the perception module along with the task graph in order to first output an updated Task State in a structured JSON format. Here, the system identifies the completed steps, in-progress steps, available steps, and immediate next step. In the second phase, the VLM uses this updated task context and multimodal data from the perception module to forecast the exact pixel coordinates  $(x, y)$  for both hands of the operator at a future time horizon. This two-phase pipeline is illustrated in Figure 3. This two-phase approach is crucial because it decouples long-horizon semantic reasoning about the overall task from short-horizon spatiotemporal forecasting of immediate motion, allowing each phase to be optimized with the most relevant context. In this module, various prompting strategies are investigated to optimize the VLM's predictive accuracy, which are discussed in Section 5.



**Figure 3.** The two-phase pipeline of the Reasoning and Intent Prediction Module. (a) In the first phase, the VLM uses the task graph and visual information to update the overall task state; (b) in the second phase, it uses this state along with a sequence of recent frames and additional context to predict the operator's future hand positions.

### 3.3. Human Motion Generation

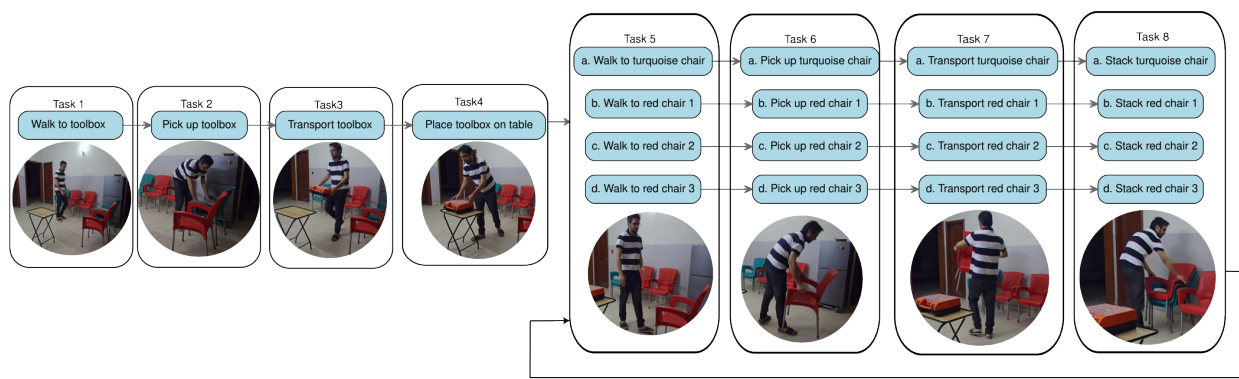
As a final, downstream step, the high-level predictions from the Reasoning and Intent Prediction Module are used to drive the Motion Generation Module. This module is crucial for visualizing the predicted intent as a full-body 3D motion. The core of this module adapts the CLoSD framework [35], which builds upon the Human Motion Diffusion Model (MDM) [32]. The CLoSD framework utilizes a forward diffusion process  $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1 - \alpha_t)I)$  that gradually adds noise to a motion sequence, while the reverse process  $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I)$  learns to denoise it by minimizing an L2 loss  $\mathcal{L}_{\text{simple}} = \mathbb{E}_{x_0, t} [\|x_0 - \hat{x}_0\|_2^2]$ . The model receives the previous 40 frames of motion data in HumanML3D format and synthesizes the future motion sequences for the next two seconds (60 frames). This is additionally conditioned on a textual prompt and a target joint (e.g., `right_wrist`, `pelvis`) and its target 3D location. Hence, the Human Motion Model synthesizes a physically plausible and semantically guided future motion sequence. This motion sequence is subsequently visualized in a virtual representation of the real environment provided by a 3D point cloud created by the perception module's depth estimation model.

## 4. Framework Validation and Use Cases

This section validates the proposed framework by applying it to three diverse use cases. Each scenario is designed to test the framework's predictive capabilities in a different context of varying difficulty.

### 4.1. Use Case 1: Chair Stacking Scenario

This use case introduces a predictable task to demonstrate the framework's core capabilities: a person moving a toolbox from a chair and then stacking several chairs in a domestic environment, as shown in Figure 4.



**Figure 4.** Use Case 1: Stacking task in a domestic setting.

The initial phase involves human–AI collaboration to define the task context. An AI model analyzes a sample video of the task or textual instructions to propose an initial “Task Context Description”. This description includes the overall goal (e.g., “stack all chairs in the designated area”), a state representation (e.g., operator holding status, ordered list of stacked chairs), a decomposition of actions (e.g., “walk to turquoise chair”, “pick up red chair”, “transport turquoise chair”), and their dependencies. This draft is then reviewed and refined by a human. The LLM’s output is a JSON object representing the updated state variables based on the task graph, including *gaze\_target*, *operator\_holding*, *num\_chairs\_stacked*, *steps\_completed*, and *steps\_available*.

In the first phase, the overall task state is predicted by the framework. When the overall task state is understood, the framework can proceed to a more granular short-term motion prediction task, as illustrated in Figure 5. For this, the VLM is provided with a rich multimodal context. This includes a sequence of the ten most recent image frames, captured at intervals of 0.2 s. Each frame is augmented with vital information: a heatmap indicating the operator’s gaze target, an overlay of the estimated human pose, and the precise normalized pixel coordinates and velocities of the operator’s hands. The model is also provided with the current sub-task description in text form (e.g., “transport turquoise chair”). The VLM’s objective is to analyze this temporal sequence, then, informed by the task context, to predict the exact pixel locations of the left and right hands two seconds after the final frame in the sequence. This approach moves beyond simple kinematic extrapolation, compelling the model to make a physically and semantically plausible forecast based on an understanding of the human’s immediate goal.

Depending on the approach, the  $x$ ,  $y$  pixel coordinates or the target object can be processed to find the world-frame coordinates of the target joint, as described in Section 3.1. This is further employed in the motion generation module. Figure 6 illustrates the motion generation results for the chair-stacking use case while performing Task 4: “place toolbox on chair”.

The modularity of the framework also allows for flexible integration of alternative motion visualization methods. One such alternative is the use of AI driven image-to-video generation tools. In this approach, the reasoning and intent prediction module of the framework generates a text prompt describing the predicted intent. The prompt and the starting frame of the sequence are passed to a video generation model. This method is not suitable for real-time prediction due to significant computational latency. Additionally, it can sometimes lead to physically implausible hallucinations. However, it serves as a powerful tool for visualization. Figure 7 illustrates a comparison between actual motion frames and those synthesized by the Wan2.1 framework [48] using this technique.

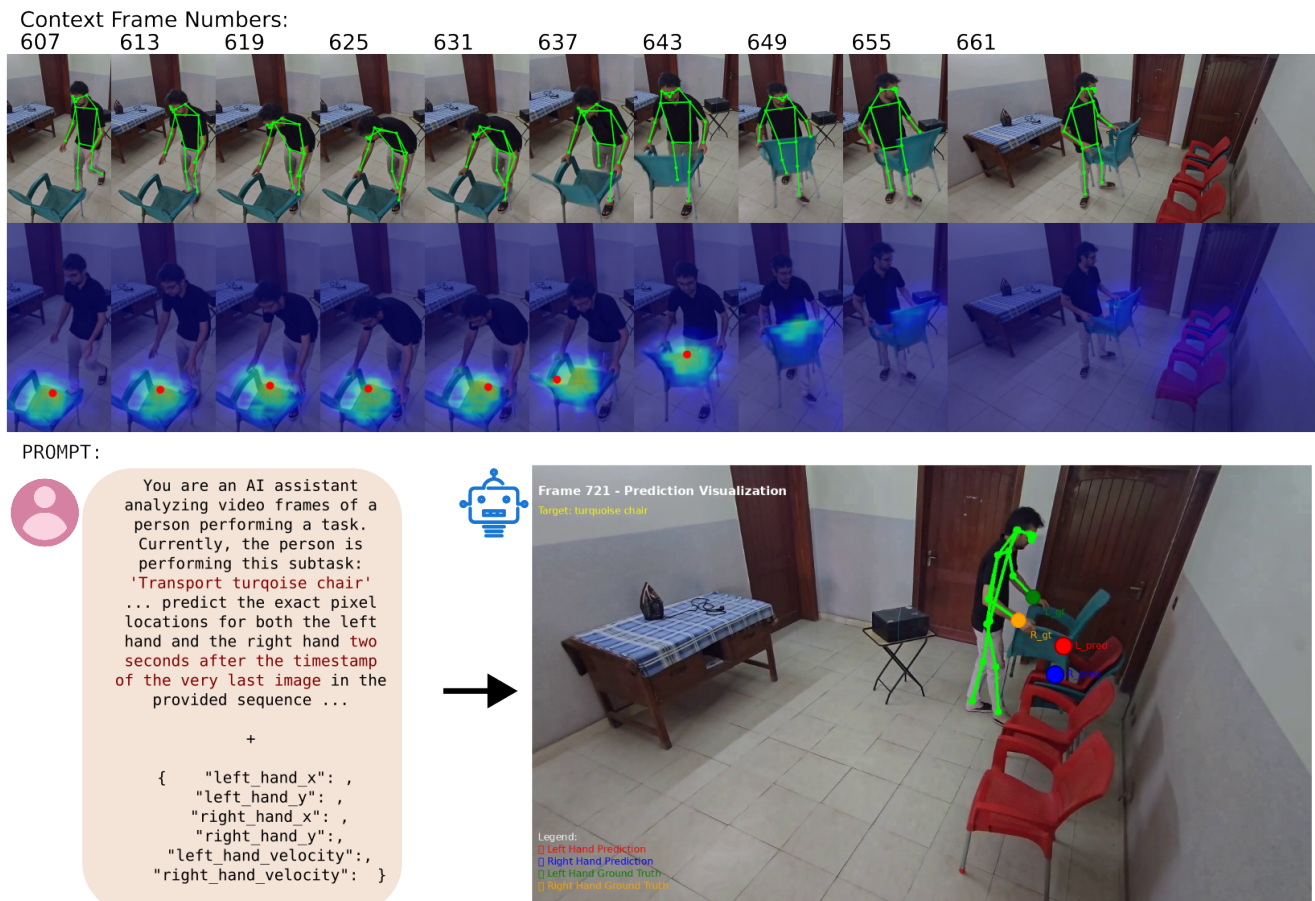


Figure 5. Hand position prediction for Stacking scenario.



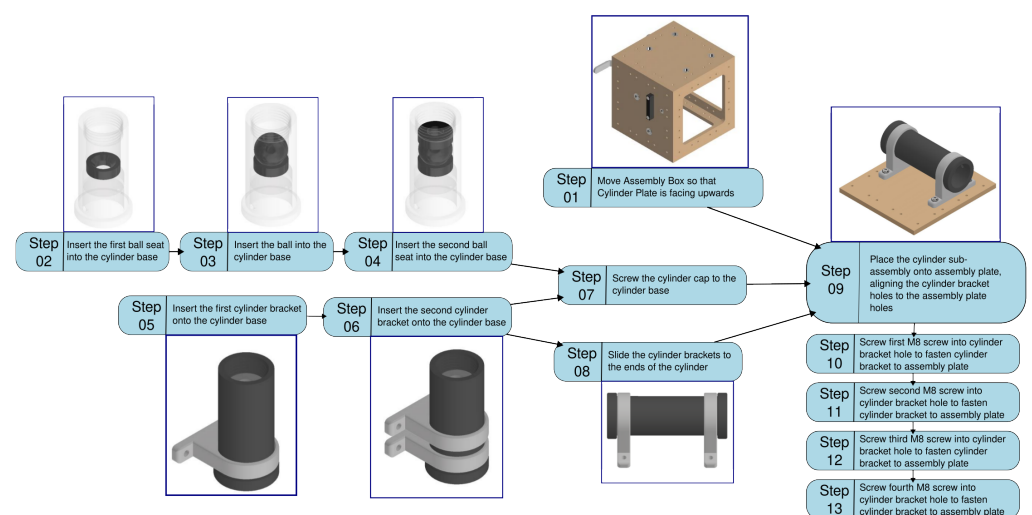
Figure 6. Visualizations for Task 4 within the Stacking use case. (a) Starting position (in red) and ground truth of the human's motion (in green) as they place the toolbox on the table; (b) generated motion (in blue), with the yellow line representing the target vector for the position of the left wrist.



**Figure 7.** Comparison of actual and AI-generated motion frames for placing a toolbox: (a) actual motion frames and (b) AI-generated motion frames.

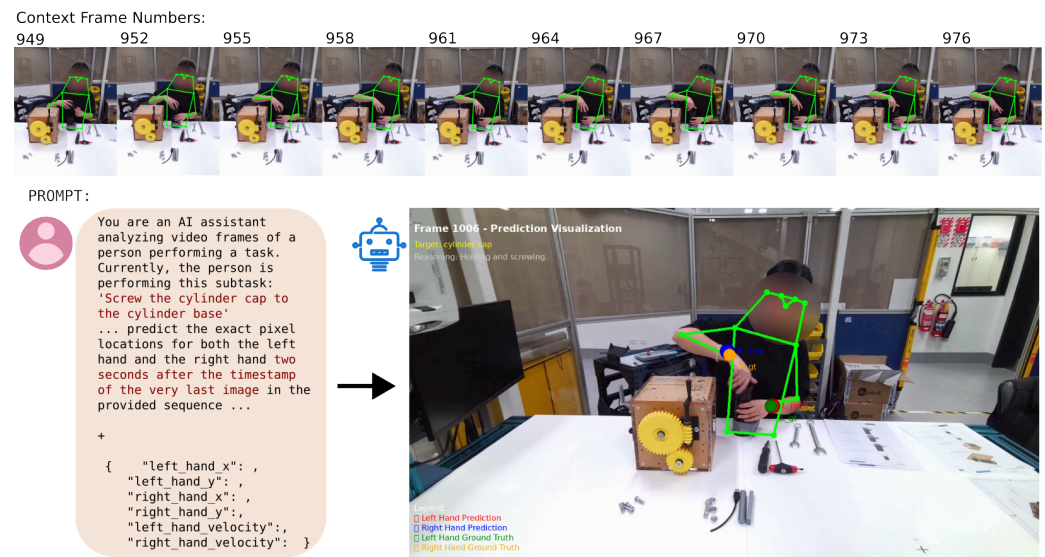
#### 4.2. Use Case 2: Assembly Task

The second use case leverages the Human Assembly Video Dataset (HaVID) [18] to evaluate the framework’s performance on fine-grained industrial assembly tasks. HaVID’s detailed annotations of actions provide a rich ground truth for assessing intent prediction. This use case demonstrates the framework’s applicability to tasks with clear SOPs, common in manufacturing environments [49]. PDF assembly instructions and exploded assembly drawings for a task such as HaVID’s “Cylinder Assembly” are first used with AI assistance to generate a comprehensive action graph (DAG) and a suitable state representation (Figure 8). Given the potential for variations in assembly sequences, multiple valid paths through the DAG are considered.



**Figure 8.** Directed Action Graph (DAG) for Assembly Task based on the HAViD dataset.

Ground truth states for selected HaVID video segments are then generated by translating HaVID's temporal annotations into the state representation format in a process facilitated by AI with human oversight. These ground truth states serve for in-context learning examples as well as for evaluation. As detailed in the previous use case, the system first predicts the system state. This is followed by a fine-grained prediction of hand positions to forecast the operator's immediate movements, applying the same multimodal analysis approach (Figure 9).



**Figure 9.** Hand position prediction for Assembly scenario.

The results of this intent-aware forecasting are then used to generate plausible future motion, as visualized in Figure 10.



**Figure 10.** Visualizations within the HAVID Assembly use case. (a) Ground truth of the human's motion over the previous 1 s (in green) and (b) generated motion (in blue), indicating the expected motion of the human's hand.

#### 4.3. Use Case 3: Cooking Scenario

To test the framework against long-horizon and complex tasks, the third use case utilizes the EgoExo4D dataset [19]. This dataset is particularly suitable for this research because it includes synchronized egocentric and exocentric video streams along with rich temporal annotations, including task and keystep-level labels. A challenging noodle cooking scenario was selected for which multiple recordings from the same kitchen location were available, providing ideal in-context learning examples.

The cooking task (Figure 11) introduces distinct challenges that are not as prevalent in the other scenarios. It involves a much larger time horizon, and the state of the task

is often ambiguous and not easily discernible from visual cues alone. For instance, by observing a single frame or even a short sequence of an idle cook, it is difficult to determine if the vegetables are fully chopped or if the noodles have finished boiling; furthermore, the scenario includes significant periods of inactivity or “waiting” during which the operator’s next action is not imminent. While the overall task has a clear structure, the cook does not need to move from completing one sub-task to immediately carrying out the next, making the precise timing of future actions inherently unpredictable.

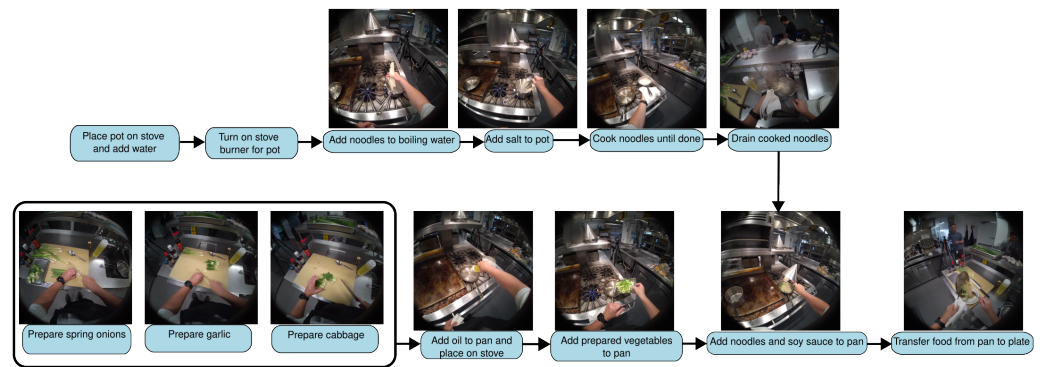


Figure 11. Sample action graph for Cooking scenario.

Despite these complexities, the framework follows the established two-phase prediction process, with the AI agent first predicting the current action or task state, then using this to inform a fine-grained hand position prediction. A key distinction in this use case is the augmentation of the VLM’s context with first-person egocentric frames, which are provided to the model along with the primary exocentric view and gaze data, as shown in Figure 12. For brevity, only half the context frames are shown in the figure. The RGB frames provided as context are also omitted.

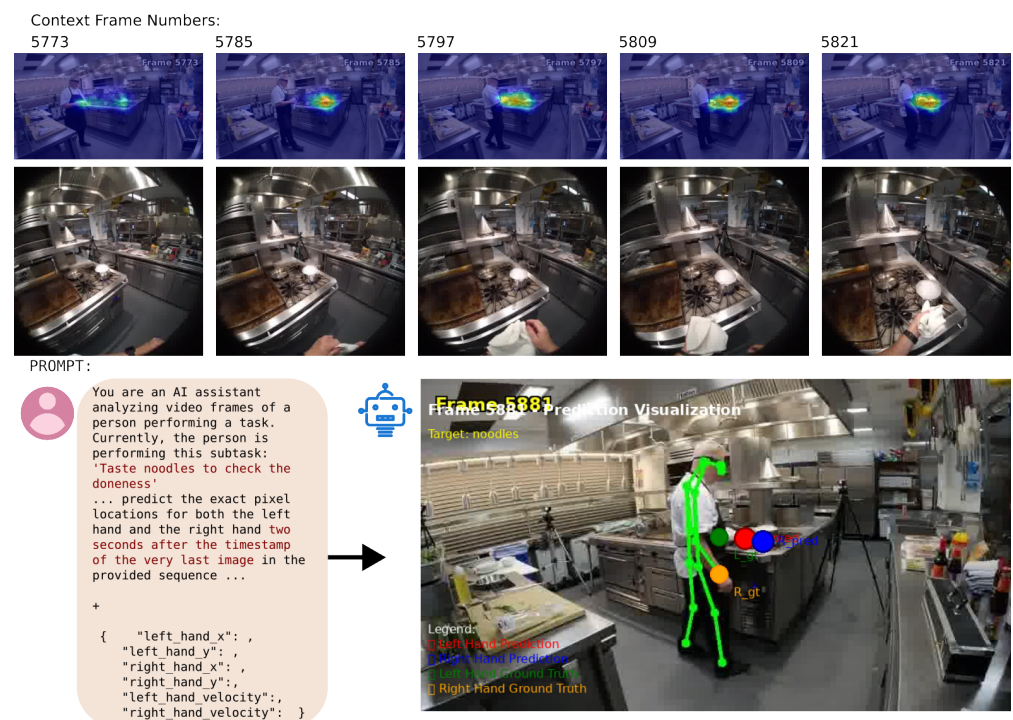


Figure 12. Hand position prediction for Cooking scenario.

## 5. Framework Performance Evaluation

Evaluating the performance of the proposed framework requires a multi-faceted approach, as no single established benchmark currently exists for the end-to-end task of context-aware human intention and motion prediction. Our work aims to lay the groundwork for such a benchmark; therefore, we define a set of specialized metrics to assess both the high-level semantic understanding (Phase 1) and low-level motion prediction accuracy (Phase 2), and report results across diverse use cases.

To evaluate the frame-wise prediction of the task state, we use a weighted F1-score, which is well-suited for this multi-label classification problem [50]:

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} = \frac{2TP}{2TP + FP + FN} \quad (2)$$

where  $TP$ ,  $FP$ , and  $FN$  are the counts of true positives, false positives, and false negatives, respectively, and  $P$  and  $R$  are the precision and recall, respectively. The overall Task State Accuracy (TSA) is calculated as a weighted sum of the F1-scores for the classification of completed steps, in-progress steps, available steps, and the immediate next step, as follows:

$$\text{TSA Score} = (w_c F1_c + w_p F1_p + w_a F1_a + w_n F1_n) \quad (3)$$

where  $F1_c$ ,  $F1_p$ ,  $F1_a$ , and  $F1_n$  are the F1-scores for the completed, in-progress, available, and immediate next steps, respectively, with weights  $w_c$ ,  $w_p$ ,  $w_a$ , and  $w_n$ .

Table 1 presents performance figures for the Task State Accuracy. A central observation is the tradeoff between accuracy, latency, and cost. A baseline strategy of providing only a single image to the Vision–Language Model (VLM) consistently underperforms due to the lack of historical context or “memory”. For a balance of performance and latency, a Rolling Context Window (RCW) strategy was employed using ten frames spaced one second apart, augmented with the predicted state from the initial frame. However, this approach is prone to consistency bias; if the model erroneously determines that a step has been completed, it often fails to revise this belief, causing subsequent predictions to suffer. When this strategy is tested with the ground-truth state provided as context, performance is significantly higher and becomes less dependent on model size, with smaller models also performing well. This approach can achieve near real-time predictions (within 2–3 s) when using the low-latency Gemini 2.5 Flash-Lite model, whereas Gemma and the standard Gemini 2.5 Flash models average closer to 10 s. The best accuracy is achieved using in-context learning, where a similar full example video and corresponding evolution of its ground-truth state are provided in the prompt. While effective, the associated latency makes this approach impractical for real-time applications with current technology.

The inclusion of additional visual cues yielded mixed results. A significant improvement was observed by incorporating egocentric first-person views in the EgoExo4D-based Cooking scenario, where performance increased from 0.568 to 0.629, a 10.7% improvement. In contrast, adding gaze target heatmaps, when averaged over all experiments, led to only an inconsistent and marginal 3% performance increase, from 0.667 to 0.688. We theorize that the egocentric perspective provides a more reliable signal of operator attention than gaze target heatmaps. In addition, this suggests that while gaze data are valuable, their primary utility may be in predicting short-term hand motion, as discussed later, rather than long-horizon task state analysis.

**Table 1.** Aggregated performance scores and rankings for Task State Accuracy.

Exp. Group	Model	Task-Specific Score			Performance	
		Stack	Assembly	Cooking	Overall	Rank
Single Image	Gemini 2.5 Flash Lite	0.557	0.367	0.285	0.403	
RCW with Ground Truth	Gemini 2.5 Pro	0.784	0.794	0.726	0.768	1
	Gemini 2.5 Flash	0.763	0.781	0.721	0.755	2
	Gemini 2.5 Flash Lite	0.703	0.724	0.813	0.747	3
	Gemma-27B	0.674	0.776	0.689	0.713	4
In-Context Learning	Gemini 2.5 Flash	0.793	0.634	0.703	0.710	1
	Gemini 2.5 Pro	0.850	0.689	0.578	0.706	2
	Gemini 2.5 Flash Lite	0.703	0.637	0.571	0.637	3
RCW with Predicted State	Gemini 2.5 Pro	0.780	0.753	0.520	0.684	1
	Gemini 2.5 Flash	0.735	0.679	0.535	0.650	2
	Gemini 2.5 Flash Lite	0.695	0.642	0.567	0.635	3
	Gemma-27B	0.634	0.659	0.478	0.590	4

An effective strategy for LLMs should achieve high performance with minimal computational cost. Figure 13 illustrates the tradeoff between performance and the average number of tokens generated, which is a proxy for computational expense.

While providing more context can enrich a model’s understanding, this approach is not without drawbacks. In addition to the obvious increases in latency and cost, simply allocating more test-time compute by expanding the context can be actively detrimental to performance. Recent work has highlighted the phenomenon of inverse scaling in test-time compute, where model performance paradoxically decreases as they are provided with more computational resources to generate a response [51]. This suggests that models can “overthink” or become lost in an unnecessarily large context, leading to performance degradation. The ICL strategy, with its massive token count, runs a higher risk of encountering this issue. By being more selective with the provided context, RCW methods not only reduce latency and costs but also mitigate the risk of inverse scaling, demonstrating a more robust and efficient approach.

The temporal aspect of the predictions is critical. We plotted the timeliness of the model’s determination that a step was completed for each use case. The results show good agreement for the Stacking and HaVID scenarios; however, the prediction of subtask completion times for the Cooking scenario is less accurate, which is attributed to the ambiguous nature and prolonged waiting periods inherent to that task. The visualization for the best-performing models for each use case is shown in Figure 14.

For the second phase, consisting of predicting future hand positions, the Normalized Mean Position Error (NMPE) is used. This metric calculates the average L2 norm (Euclidean distance) of the pixel error between the predicted and ground-truth hand coordinates, normalized from 0 to 1000:

$$\text{NMPE} = \frac{1}{N} \sum_{i=1}^N \left( \frac{\sqrt{(x_{p,l} - x_{g,l})^2 + (y_{p,l} - y_{g,l})^2} + \sqrt{(x_{p,r} - x_{g,r})^2 + (y_{p,r} - y_{g,r})^2}}{2} \right)_i \quad (4)$$

where the subscripts  $p, g, l$ , and  $r$  respectively denote predicted, ground-truth, left hand, and right hand for each prediction instance  $i$  out of  $N$ .

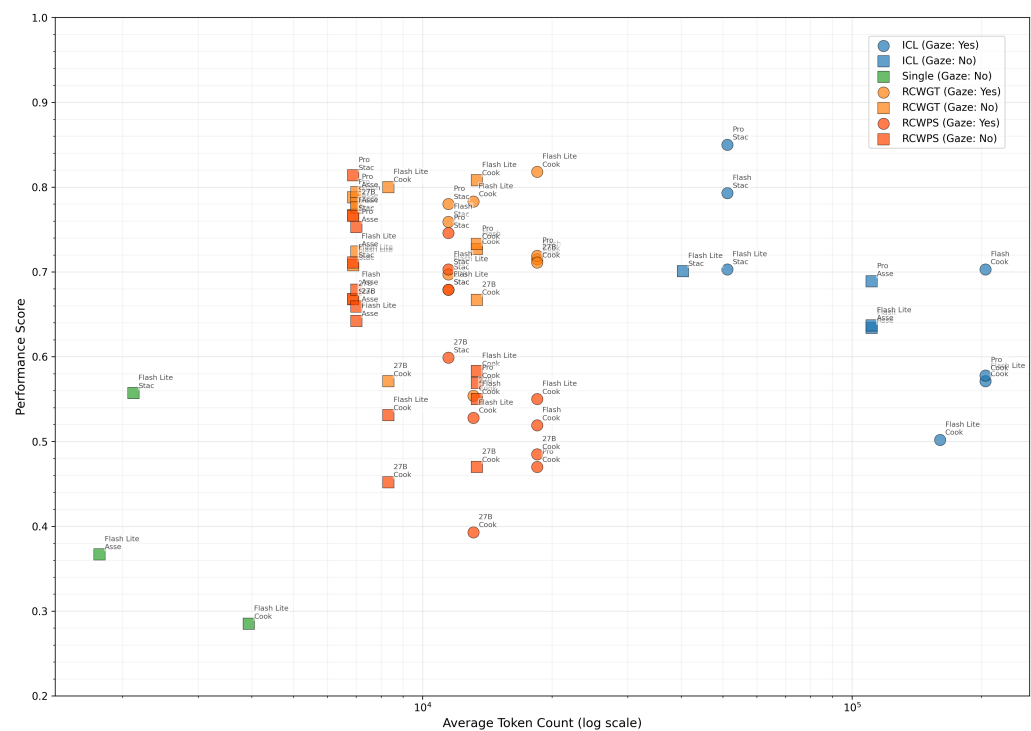


Figure 13. Performance score versus the average number of tokens per generation.

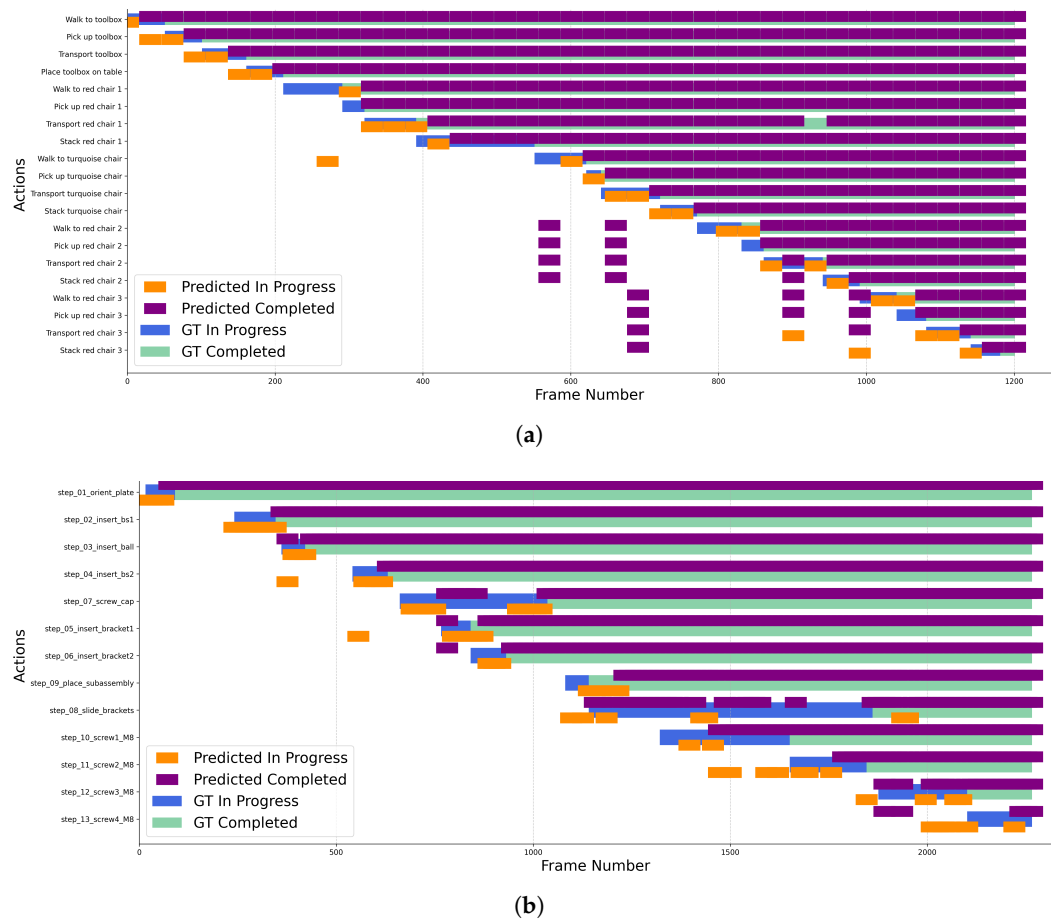
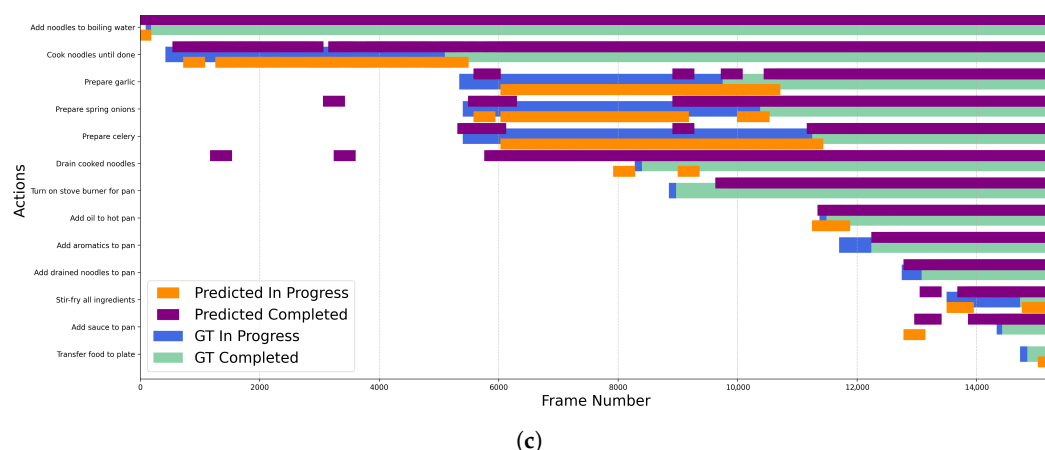


Figure 14. Cont.

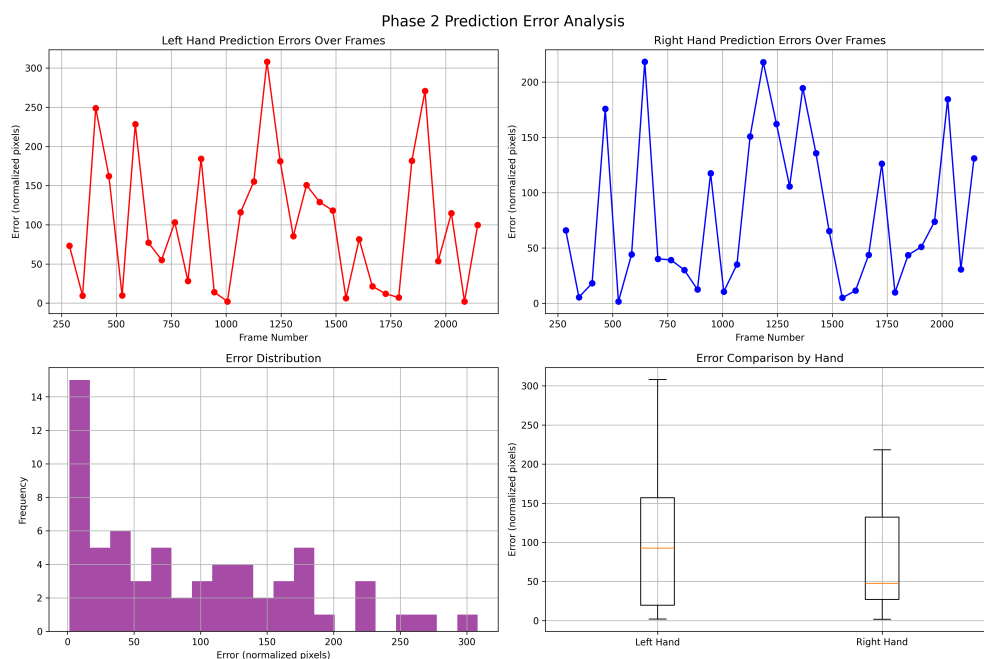


**Figure 14.** Visualization of model performance across different use cases: (a) Stacking, (b) Assembly, and (c) Cooking.

While NMPE measures accuracy, it does not capture the exploratory nature of a model's predictions. A model that simply predicts minimal movement from the last known position might achieve a low NMPE but fail to anticipate significant goal-directed actions. To quantify this, a Prediction Diversity metric is introduced. This metric measures how far a prediction deviates from the recent trajectory, rewarding predictions that are not mere extrapolations. It is calculated as the average of the Euclidean distances from the predicted point to both the average position and the final position of the hand in the input sequence, normalized by the standard deviation of the input positions:

$$\text{Diversity} = \frac{D(\mathbf{p}_{\text{pred}}, \bar{\mathbf{p}}_{\text{in}}) + D(\mathbf{p}_{\text{pred}}, \mathbf{p}_{\text{last}})}{2\sqrt{\sigma(\mathbf{p}_{\text{in},x})^2 + \sigma(\mathbf{p}_{\text{in},y})^2}} \quad (5)$$

where  $\mathbf{p}_{\text{pred}}$  is the predicted position,  $\bar{\mathbf{p}}_{\text{in}}$  is the average position of the input trajectory, and  $\mathbf{p}_{\text{last}}$  is the last position in the input trajectory. Figure 15 shows the detailed error analysis for a sample experiment using the Assembly scenario.



**Figure 15.** Sample Error Analysis for hand position prediction in Assembly scenario.

The performance of our two-phase pipeline is detailed in Table 2, which presents a breakdown of the Normalized Mean Position Error (NMPE) and Prediction Diversity across various use cases. The NMPE serves as our primary metric for motion prediction accuracy, while the Prediction Diversity quantifies the variability of the generated motion forecasts. A higher diversity score indicates the model’s capacity to predict significant goal-directed actions rather than simple extrapolations. Our findings show a consistent tradeoff between these two metrics, with higher prediction diversity often coming at the cost of increased position error. Additionally, we include the results of an ablation experiment using the Assembly scenario. In this “No Context” experiment, we performed hand position prediction without the task context from Phase 1. This resulted in a substantial degradation in performance, with the NMPE increasing to 202.35. This outcome underscores the necessity of our two-phase approach, which decouples long-horizon semantic reasoning from short-horizon spatiotemporal forecasting. Furthermore, our results consistently showed that few-shot predictions with one or two in-context examples provided significant accuracy improvements over a zero-shot approach.

**Table 2.** Normalized Mean Position Error (NMPE) and Prediction Diversity by use case.

Use Case	Model	Gaze Data	Examples	NMPE	Diversity
Stack	Gemini 2.5 Flash Lite	Yes	2-shot	109.46	2.02
	Gemini 2.5 Flash	Yes	2-shot	111.71	3.48
	Gemma 3 27B	Yes	0-shot	114.97	2.33
	Gemini 2.5 Flash Lite	No	2-shot	120.58	2.08
	Gemini 2.5 Flash Lite	Yes	0-shot	121.74	1.42
	Gemini 2.5 Pro	Yes	2-shot	152.68	5.68
	Gemini 2.5 Pro	Yes	1-shot	181.76	6.14
Assembly	Gemma 3 27B	N/A	1-shot	72.19	0.81
	Gemini 2.5 Flash Lite	N/A	2-shot	74.70	1.20
	Gemini 2.5 Flash Lite	N/A	0-shot	76.04	1.27
	Gemini 2.5 Flash	N/A	2-shot	81.85	3.67
	Gemini 2.5 Pro	N/A	1-shot	139.40	12.49
	Gemini 2.5 Pro (No Context)	N/A	0-shot	202.35	23.05
Cooking	Gemini 2.5 Flash Lite	Yes	0-shot	70.16	1.04
	Gemini 2.5 Flash Lite	Yes	1-shot	70.88	1.19
	Gemini 2.5 Flash Lite	No	1-shot	74.49	1.31
	Gemini 2.5 Flash	Yes	2-shot	78.36	2.69
	Gemini 2.5 Pro	Yes	1-shot	89.34	3.37

## 6. Discussion and Future Work

The proposed modular framework advances the development of proactive robotic assistants by integrating state-of-the-art Vision–Language Models (VLMs) for high-level intent prediction with advanced models for perception and motion synthesis. This approach addresses key limitations in prior work. Unlike action recognition models, which are confined to limited classes [52,53] or context-agnostic pose forecasting methods [54,55], our framework leverages the generalized reasoning capabilities of VLMs to interpret multimodal context. By engineering this context through structured task graphs and rich perceptual data, the system can infer human intent in a way that is more aligned with the complexities of real-world tasks.

Prior work in social and assistive robotics often incorporates human intention within a narrow scope, such as to assess a user’s engagement level [56] or emotional state [57] and respond appropriately. Such unstructured interactions do not involve completing a specific industrial task or adhering to formal Standard Operating Procedures (SOPs). In

contrast, our framework is designed to predict task-based intentions within structured goal-oriented workflows.

Despite its promise, the proposed framework has several limitations that can guide future research. The framework's current reliance on predefined task graphs, even when AI-assisted, restricts its applicability in completely unstructured or novel scenarios where a task plan is not available beforehand. Furthermore, the motion generation pipeline is not object-aware; while it utilizes the CLoSD model [35] for physically plausible motion synthesis, the generated motions do not explicitly account for geometric interactions with specific objects in the environment. We address this limitation pragmatically by decoupling scene understanding from motion synthesis. The VLM, which is object-aware through its analysis of visual input, predicts a target 3D coordinate for a joint (e.g., placing a hand on a specific object). This 3D point then serves as a goal for the object-agnostic motion generation module. While this approach functions as an effective workaround, a fully integrated object-aware motion model would be superior. Future work will explore integrating emerging models for physics-based human–object interaction within our modular framework. However, these methods often rely on datasets with a limited variety of objects and interactions [58–60]. Finally, the use cases are centered on a single human operator, whereas many industrial environments involve complex multi-human collaborations, which introduces additional challenges such as occlusion and the need to interpret social dynamics.

To address these limitations and expand the framework's capabilities, several avenues for future work are identified. A limitation of the current implementation is its reliance on a fixed-interval prediction cycle where motion is forecast at regular time steps regardless of the task's dynamic context. Instead, an event-driven prediction framework can be explored where inference is triggered by salient cues, such as a sudden shift in gaze or the completion of a task step.

To improve generalization, methods for dynamically generating and adapting task graphs from observation may be explored, enabling the system to reason about unfamiliar workflows. Techniques inspired by multi-agent systems research [38] could offer novel ways to coordinate the flow of information and aid decision-making. However, this approach is expected to add more latency to the system. Prompting strategies might be explored to encourage the VLM to recognize uncertainty and proactively seek clarification when its confidence is low or when critical information is missing, drawing inspiration from approaches such as [37]. Incorporating mechanisms for learning from human feedback or corrections during interaction [61] could allow the system to adapt and respond to task requirements and individual preferences. Further research could also explore the integration of additional context modalities such as audio, physiological signals, or data from IoT sensors to create a more holistic understanding of the operational environment.

Another promising future direction is to use our framework's proactive capabilities to enhance reactive impedance controllers for human-guided robots. By generating an anticipatory signal of where the user intends to move, an impedance controller can optimize its response in advance rather than reacting solely to force. This would lead to a more symbiotic physical interaction with reduced human effort and improved task performance [62]. In addition to predicting human intent for proactive assistance, the accurate forecasting of hand positions would also provide a critical safety layer for HRC. Anticipating human movement could enable robots to adjust their trajectories or speeds to avoid collisions, contributing to enhanced safety during human–robot collaboration. This capability is crucial for the adoption of heavy-payload robots in collaborative manufacturing, where safety concerns have historically been a major hurdle [63].

HCDTs powered by such human-centric situational awareness can serve as a virtual mentor for training. In addition, they can predict and flag ergonomically hazardous move-

ments to prevent workplace injuries, and provide proactive robotic support by anticipating the need for assistance or carrying out tasks in parallel. By predicting human intent within a task context, our model provides a necessary prerequisite for systems in which a robot can intelligently assist a human worker without being explicitly commanded. This capability is a key component of Industry 5.0, enabling resilient and adaptable manufacturing; furthermore, this paradigm moves us closer to prompt-based manufacturing, where high-level human intent expressed through natural language or gesture can be seamlessly translated into a machine's adaptable physical actions. This approach is essential for enabling High-Mix-Low-Volume (HMLV) production and a more human-centric industrial automation. Critically, the transition from research prototype to a practical industrial tool depends on rigorous real-world validation. Future work must involve deployment in real-time lab-based HRC scenarios, where performance is assessed not only by prediction accuracy but also by HRC-centric metrics such as task efficiency, human idle time, and operator trust. Ultimately, this could enable a transition of robots from passive collaborative robots to intelligent coworkers.

## 7. Conclusions

This paper presents a modular framework for context-aware human intent prediction and subsequent 3D motion generation, with the aim of enabling proactive robotic assistance. By integrating pretrained Vision Language Models with state-of-the-art perception and generation modules, the proposed approach provides a scalable solution that bridges high-level semantic reasoning with physically grounded motion synthesis.

Our evaluation across three use cases of varying complexity revealed critical tradeoffs between predictive accuracy, latency, and diversity. In-Context Learning (ICL) demonstrates notable limitations that counteract its potential benefits. The requirement for a large context token count results in significant computational latency, in some cases leading to a paradoxical degradation of predictive accuracy that renders the approach unsuitable for real-time applications. We identified the Rolling Context Window (RCW) strategy as a more viable approach, offering a strong balance of performance and efficiency; however, this method is susceptible to consistency bias, as an initial error in state prediction can propagate and degrade subsequent performance. A key finding was that augmenting context with egocentric video views yielded a substantial 10.7% performance increase in complex tasks.

Furthermore, we observed a tradeoff between accuracy and prediction diversity when forecasting short-term motion. More powerful models tended to generate predictions with higher diversity, attempting to forecast significant goal-directed actions rather than simple extrapolations. However, this increased diversity often came at the cost of higher position error. Additionally, predictive accuracy showed consistent improvement when providing one or two in-context examples (few-shot) compared to a zero-shot approach.

Through continued development and integration, the framework presented in this work can significantly advance the ability of robots to understand, anticipate, and effectively collaborate with humans in complex dynamic settings.

**Author Contributions:** Conceptualization, U.A. and A.K.; methodology, U.A.; software, U.A.; validation, U.A., W.A.L. and A.K.; investigation, U.A.; data curation, U.A.; writing—original draft preparation, U.A. and W.A.L.; writing—review and editing, U.A., M.M.K., A.K., W.A.L. and S.R.; visualization, U.A.; supervision, M.M.K., A.K., W.A.L. and S.R.; project administration, M.M.K., A.K. and S.R. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study. For the publicly available datasets (EgoExo4D and HA-ViD), the original data collectors were responsible for obtaining informed consent from the participants, and our study complies with the terms of use for these datasets. For the self-recorded data used in the Stacking scenario, one of the authors served as the subject and directly consented to participation.

**Data Availability Statement:** Additional materials, videos, and the open-source implementation are available at the project website: <https://usmanasad88.github.io/hcdt/> (accessed on 11 September 2025).

**Acknowledgments:** The authors would like to acknowledge the Digital Innovation Research Group in the Department of Engineering at Nottingham Trent University for their support of this work. The authors would also like to thank the developers of the open-source libraries and datasets used in this research.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AR	Augmented Reality
BT	Behavior Tree
DAG	Directed Acyclic Graph
HCDT	Human-Centric Digital Twin
HMLV	High-Mix-Low-Volume
HRC	Human-Robot Collaboration
ICL	In-Context Learning
KG	Knowledge Graph
LLM	Large Language Model
ML	Machine Learning
MoCap	Motion Capture
NMPE	Normalized Mean Position Error
PDDL	Planning Domain Definition Language
POI	Point of Interest
RCW	Rolling Context Window
SME	Small and Medium Enterprises
SMPL-X	Skinned Multi-Person Linear Model with Extensions
TSA	Task State Accuracy
VLM	Vision-Language Model
VR	Virtual Reality

## References

1. Andronas, D.; Kokotinis, G.; Makris, S. On modelling and handling of flexible materials: A review on Digital Twins and planning systems. *Procedia CIRP* **2021**, *97*, 447–452. [[CrossRef](#)]
2. Polini, W.; Corrado, A. Digital twin of composite assembly manufacturing process. *Int. J. Prod. Res.* **2020**, *58*, 5238–5252. [[CrossRef](#)]
3. Asad, U.; Khan, M.; Khalid, A.; Lughmani, W.A. Human-Centric Digital Twins in Industry: A Comprehensive Review of Enabling Technologies and Implementation Strategies. *Sensors* **2023**, *23*, 3938. [[CrossRef](#)]
4. Umbrico, A.; Orlandini, A.; Cesta, A.; Faroni, M.; Beschi, M.; Pedrocchi, N.; Scala, A.; Tavormina, P.; Koukas, S.; Zalonis, A.; et al. Design of Advanced Human-Robot Collaborative Cells for Personalized Human-Robot Collaborations. *Appl. Sci.* **2022**, *12*, 6839. [[CrossRef](#)]
5. Cotta, W.A.A.; Lopes, S.I.; Vassallo, R.F. Towards the Cognitive Factory in Industry 5.0: From Concept to Implementation. *Smart Cities* **2023**, *6*, 1901–1921. [[CrossRef](#)]

6. Sirintuna, D.; Kastritsi, T.; Ozdamar, I.; Gandarias, J.M.; Ajoudani, A. Enhancing human-robot collaborative transportation through obstacle-aware vibrotactile warning and virtual fixtures. *Robot. Auton. Syst.* **2024**, *178*, 104725. [\[CrossRef\]](#)
7. Nezhad, H.Y.; Wang, X.; Court, S.D.; Thapa, B.; Erkoyuncu, J.A. Development of an augmented reality equipped composites bonded assembly and repair for aerospace applications. *IFAC-PapersOnLine* **2020**, *53*, 209–215. [\[CrossRef\]](#)
8. Laughlin, B.D.; Skelton, M.M. Augmented Reality System for Manufacturing Composite Parts. US Patent 16/167,636, 23 April 2020.
9. Mandlekar, A.; Xu, D.; Wong, J.; Nasiriany, S.; Wang, C.; Kulkarni, R.; Fei-Fei, L.; Savarese, S.; Zhu, Y.; Martín-Martín, R. What Matters in Learning from Offline Human Demonstrations for Robot Manipulation. *arXiv* **2021**, arXiv:2108.03298. [\[CrossRef\]](#)
10. Chi, C.; Feng, S.; Du, Y.; Xu, Z.; Cousineau, E.; Burchfiel, B.; Song, S. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. *arXiv* **2023**, arXiv:2303.04137.
11. Yu, Q.; Moghani, M.; Dharmarajan, K.; Schorp, V.; Panitch, W.C.H.; Liu, J.; Hari, K.; Huang, H.; Mittal, M.; Goldberg, K.; et al. ORBIT-Surgical: An Open-Simulation Framework for Learning Surgical Augmented Dexterity. *arXiv* **2024**, arXiv:2404.16027.
12. Liu, F.; Su, E.; Lu, J.; Li, M.; Yip, M.C. Robotic Manipulation of Deformable Rope-Like Objects Using Differentiable Compliant Position-Based Dynamics. *IEEE Robot. Autom. Lett.* **2023**, *8*, 3964–3971. [\[CrossRef\]](#)
13. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv* **2022**, arXiv:2201.11903.
14. Fan, H.; Liu, X.; Fuh, J.Y.H.; Lu, W.F.; Li, B. Embodied intelligence in manufacturing: Leveraging large language models for autonomous industrial robotics. *J. Intell. Manuf.* **2024**, *36*, 1141–1157. [\[CrossRef\]](#)
15. Li, F.; Zhang, R.; Zhang, H.; Zhang, Y.; Li, B.; Li, W.; Ma, Z.; Li, C. LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. *arXiv* **2024**, arXiv:2407.07895.
16. Singh, I.; Blukis, V.; Mousavian, A.; Goyal, A.; Xu, D.; Tremblay, J.; Fox, D.; Thomason, J.; Garg, A. ProgPrompt: Program generation for situated robot task planning using large language models. *Auton. Robot.* **2023**, *47*, 999–1012. [\[CrossRef\]](#)
17. Ahn, M.; Brohan, A.; Brown, N.; Chebotar, Y.; Cortes, O.; David, B.; Finn, C.; Fu, C.; Gopalakrishnan, K.; Hausman, K.; et al. Do as I Can, Not as I Say: Grounding Language in Robotic Affordances. *arXiv* **2022**, arXiv:2204.01691. [\[CrossRef\]](#)
18. Zheng, H.; Lee, R.; Lu, Y. HA-ViD: A human assembly video dataset for comprehensive assembly knowledge understanding. In Proceedings of the 37th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 10–16 December 2023.
19. Grauman, K.; Westbury, A.; Torresani, L.; Kitani, K.; Malik, J.; Afouras, T.; Ashutosh, K.; Baiyya, V.; Bansal, S.; Boote, B.; et al. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 19383–19400. [\[CrossRef\]](#)
20. Huang, Z.; Hasan, A.; Shin, K.; Li, R.; Driggs-Campbell, K. Long-Term Pedestrian Trajectory Prediction Using Mutable Intention Filter and Warp LSTM. *IEEE Robot. Autom. Lett.* **2021**, *6*, 542–549. [\[CrossRef\]](#)
21. Huang, Z.; Mun, Y.J.; Li, X.; Xie, Y.; Zhong, N.; Liang, W.; Geng, J.; Chen, T.; Driggs-Campbell, K. Hierarchical Intention Tracking for Robust Human-Robot Collaboration in Industrial Assembly Tasks. *arXiv* **2022**, arXiv:2203.09063.
22. Mangin, O.; Roncone, A.; Scassellati, B. How to be Helpful? Supportive Behaviors and Personalization for Human-Robot Collaboration. *Front. Robot. AI* **2022**, *8*, 725780. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Zhong, R.; Hu, B.; Hong, Z.; Zhang, Z.; Lou, S.; Song, X.; Feng, Y.; Tan, J. Human-Robot handover task intention recognition framework by fusing human digital twin and deep domain adaptation. *J. Eng. Des.* **2024**, 1–17 [\[CrossRef\]](#)
24. Ding, P.; Zhang, J.; Zheng, P.; Zhang, P.; Fei, B.; Xu, Z. Dynamic scenario-enhanced diverse human motion prediction network for proactive human-robot collaboration in customized assembly tasks. *J. Intell. Manuf.* **2025**, *36*, 4593–4612. [\[CrossRef\]](#)
25. Mascaro, E.V.; Ahn, H.; Lee, D. Intention-Conditioned Long-Term Human Egocentric Action Forecasting. *arXiv* **2022**, arXiv:2207.12080.
26. Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; et al. Ego4D: Around the World in 3,000 Hours of Egocentric Video. *arXiv* **2021**, arXiv:2110.07058. [\[CrossRef\]](#)
27. Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; Cheng, L. Generating Diverse and Natural 3D Human Motions from Text. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5152–5161.
28. Peng, X.B.; Ma, Z.; Abbeel, P.; Levine, S.; Kanazawa, A. AMP: Adversarial Motion Priors for Stylized Physics-Based Character Control. *ACM Trans. Graph. (TOG)* **2021**, *40*, 1–20. [\[CrossRef\]](#)
29. Escontrela, A.; Peng, X.B.; Yu, W.; Zhang, T.; Iscen, A.; Goldberg, K.; Abbeel, P. Adversarial Motion Priors Make Good Substitutes for Complex Reward Functions. In Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 23–27 October 2022.
30. Peng, X.B.; Guo, Y.; Halper, L.; Levine, S.; Fidler, S. ASE: Large-scale Reusable Adversarial Skill Embeddings for Physically Simulated Characters. *ACM Trans. Graph. (TOG)* **2022**, *41*, 1–17. [\[CrossRef\]](#)

31. Tessler, C.; Kasten, Y.; Guo, Y.; Mannor, S.; Chechik, G.; Peng, X.B. CALM: Conditional Adversarial Latent Models for Directable Virtual Characters. In Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings, Los Angeles, CA, USA, 6–10 August 2023; pp. 1–9. [\[CrossRef\]](#)
32. Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-Or, D.; Bermano, A.H. Human Motion Diffusion Model. In Proceedings of the The Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
33. Yuan, Y.; Song, J.; Iqbal, U.; Vahdat, A.; Kautz, J. PhysDiff: Physics-Guided Human Motion Diffusion Model. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 15964–15975. [\[CrossRef\]](#)
34. Tessler, C.; Guo, Y.; Nabati, O.; Chechik, G.; Peng, X.B. MaskedMimic: Unified Physics-Based Character Control Through Masked Motion. *ACM Trans. Graph. (TOG)* **2024**, *43*, 1–21. [\[CrossRef\]](#)
35. Tevet, G.; Raab, S.; Cohan, S.; Reda, D.; Luo, Z.; Peng, X.B.; Bermano, A.H.; van de Panne, M. CLoSD: Closing the Loop between Simulation and Diffusion for multi-task character control. *arXiv* **2024**, arXiv:2410.03441.
36. Ha, H.; Florence, P.; Song, S. Scaling Up and Distilling Down: Language-Guided Robot Skill Acquisition. *arXiv* **2023**, arXiv:2307.14535. [\[CrossRef\]](#)
37. Chen, X.; Zhang, S.; Zhang, P.; Zhao, L.; Chen, J. Asking Before Action: Gather Information in Embodied Decision Making with Language Models. *arXiv* **2023**, arXiv:2305.15695.
38. Fourney, A.; Bansal, G.; Mozannar, H.; Tan, C.; Salinas, E.; Gerrits, J.; Alber, J.; Niedtner, F.; Proebsting, G.; Bassman, G.; et al. Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks. *arXiv* **2024**, arXiv:2411.04468. [\[CrossRef\]](#)
39. Mahmood, N.; Ghorbani, N.; Troje, N.F.; Pons-Moll, G.; Black, M.J. AMASS: Archive of Motion Capture as Surface Shapes. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5442–5451.
40. Plappert, M.; Mandery, C.; Asfour, T. The KIT Motion-Language Dataset. *Big Data* **2016**, *4*, 236–252. [\[CrossRef\]](#) [\[PubMed\]](#)
41. Sener, F.; Chatterjee, D.; Shelepov, D.; He, K.; Singhania, D.; Wang, R.; Yao, A. Assembly101: A Large-Scale Multi-View Video Dataset for Understanding Procedural Activities. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
42. Baradel, F.; Armando, M.; Galaaoui, S.; Brégier, R.; Weinzaepfel, P.; Rogez, G.; Lucas, T. Multi-HMR: Multi-Person Whole-Body Human Mesh Recovery in a Single Shot. In Proceedings of the 2024 European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024.
43. Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A.A.A.; Tzionas, D.; Black, M.J. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 10975–10985.
44. Shen, Z.; Pi, H.; Xia, Y.; Cen, Z.; Peng, S.; Hu, Z.; Bao, H.; Hu, R.; Zhou, X. World-Grounded Human Motion Recovery via Gravity-View Coordinates. In Proceedings of the SIGGRAPH Asia Conference Proceedings, Tokyo, Japan, 3–6 December 2024.
45. Piccinelli, L.; Sakaridis, C.; Segu, M.; Yang, Y.H.; Li, S.; Abbeloos, W.; Van Gool, L. UniK3D: Universal Camera Monocular 3D Estimation. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025.
46. Ryan, F.; Bati, A.; Lee, S.; Bolya, D.; Hoffman, J.; Rehg, J.M. Gaze-LLE: Gaze Target Estimation via Large-Scale Learned Encoders. *arXiv* **2024**, arXiv:2412.09586.
47. Tuli, T.B.; Kohl, L.; Chala, S.A.; Manns, M.; Ansari, F. Knowledge-Based Digital Twin for Predicting Interactions in Human-Robot Collaboration. In Proceedings of the 2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), Vasteras, Sweden, 7–10 September 2021; pp. 1–8. [\[CrossRef\]](#)
48. Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; et al. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv* **2025**, arXiv:2503.20314. [\[CrossRef\]](#)
49. David, J.; Coatanéa, E.; Lobov, A. Deploying OWL ontologies for semantic mediation of mixed-reality interactions for human–robot collaborative assembly. *J. Manuf. Syst.* **2023**, *70*, 359–381. [\[CrossRef\]](#)
50. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008.
51. Gema, A.P.; Hägele, A.; Chen, R.; Arditi, A.; Goldman-Wetzler, J.; Fraser-Taliente, K.; Sleight, H.; Petrini, L.; Michael, J.; Alex, B.; et al. Inverse Scaling in Test-Time Compute. *arXiv* **2025**, arXiv:2507.14417. [\[CrossRef\]](#)
52. Gong, D.; Lee, J.; Kim, M.; Ha, S.J.; Cho, M. Future Transformer for Long-term Action Anticipation. *arXiv* **2022**, arXiv:2205.14022. [\[CrossRef\]](#)
53. Chavis, Z.; Guy, S.J.; Park, H.S. Improving Keystep Recognition in Ego-Video via Dexterous Focus. *arXiv* **2025**, arXiv:2506.00827. [\[CrossRef\]](#)
54. Sofianos, T.; Sampieri, A.; Franco, L.; Galasso, F. Space-Time-Separable Graph Convolutional Network for Pose Forecasting. *arXiv* **2021**, arXiv:2110.04573.

55. Tian, S.; Zheng, M.; Liang, X. TransFusion: A practical and effective transformer-based diffusion model for 3d human motion prediction. *IEEE Robot. Autom. Lett.* **2024**, *9*, 6232–6239. [[CrossRef](#)]
56. Bi, J.; Hu, F.c.; Wang, Y.j.; Luo, M.n.; He, M. A method based on interpretable machine learning for recognizing the intensity of human engagement intention. *Sci. Rep.* **2023**, *13*, 2537. [[CrossRef](#)] [[PubMed](#)]
57. Scassellati, B. Theory of Mind for a Humanoid Robot. *Auton. Robot.* **2002**, *12*, 13–24. [[CrossRef](#)]
58. Xu, S.; Ling, H.Y.; Wang, Y.X.; Gui, L.Y. InterMimic: Towards Universal Whole-Body Control for Physics-Based Human-Object Interactions. In Proceedings of the 2025 IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR), Nashville, TN, USA, 11–15 June 2025.
59. Banerjee, P.; Shkodrani, S.; Moulon, P.; Hampali, S.; Han, S.; Zhang, F.; Zhang, L.; Fountain, J.; Miller, E.; Basol, S.; et al. HOT3D: Hand and Object Tracking in 3D from Egocentric Multi-View Videos. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025.
60. Fan, Z.; Taheri, O.; Tzionas, D.; Kocabas, M.; Kaufmann, M.; Black, M.J.; Hilliges, O. ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023.
61. Abeyruwan, S.; Graesser, L.; D’Ambrosio, D.B.; Singh, A.; Shankar, A.; Bewley, A.; Jain, D.; Choromanski, K.; Sanketi, P.R. i-Sim2Real: Reinforcement Learning of Robotic Policies in Tight Human-Robot Interaction Loops. *arXiv* **2022**, arXiv:2207.06572.
62. Xing, X.; Burdet, E.; Si, W.; Yang, C.; Li, Y. Impedance Learning for Human-Guided Robots in Contact with Unknown Environments. *IEEE Trans. Robot.* **2023**, *39*, 3705–3721. [[CrossRef](#)]
63. Asad, U.; Rasheed, S.; Lughmani, W.A.; Kazim, T.; Khalid, A.; Pannek, J. Biomechanical Modeling of Human–Robot Accident Scenarios: A Computational Assessment for Heavy-Payload-Capacity Robots. *Appl. Sci.* **2023**, *13*, 1957. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.