electronics

Article

Generalizable Interaction Recognition for Learning from Demonstration Using Wrist and Object Trajectories

Jagannatha Charjee Pyaraka [1,*], Mats Isaksson [1], John McCormick [2], Sheila Sutjipto [3] and Fouad Sukkar [3]

[1] Department of Mechanical Engineering and Product Design Engineering, School of Engineering, Swinburne University, Hawthorn Campus, Melbourne, VIC 3122, Australia; misaksson@swin.edu.au
[2] Centre for Transformative Media Technologies, Swinburne University of Technology, Hawthorn Campus, Melbourne, VIC 3122, Australia; jmccormick@swin.edu.au
[3] UTS Robotics Institute, School for Mechanical and Mechatronic Engineering, University of Technology Sydney, Ultimo, NSW 2007, Australia; sheila.sutjipto@uts.edu.au (S.S.); fouad.sukkar@uts.edu.au (F.S.)
* Correspondence: jagannathacharjeepya@swin.edu.au

Abstract

Learning from Demonstration (LfD) enables robots to acquire manipulation skills by observing human actions. However, existing methods often face challenges such as high computational cost, limited generalizability, and a loss of key interaction details. This study presents a compact representation for interaction recognition in LfD that encodes human–object interactions using 2D wrist trajectories and 3D object poses. A lightweight extraction pipeline combines MediaPipe-based wrist tracking with FoundationPose-based 6-DoF object estimation to obtain these trajectories directly from RGB-D video without specialized sensors or heavy preprocessing. Experiments on the GRAB and FPHA datasets show that the representation effectively captures task-relevant interactions, achieving 94.6% accuracy on GRAB and 96.0% on FPHA with well-calibrated probability predictions. Both Bidirectional Long Short-Term Memory (Bi-LSTM) with attention and Transformer architectures deliver consistent performance, confirming robustness and generalizability. The method achieves sub-second inference, a memory footprint under 1 GB, and reliable operation on both GPU and CPU platforms, enabling deployment on edge devices such as NVIDIA Jetson. By bridging pose-based and object-centric paradigms, this approach offers a compact and efficient foundation for scalable robot learning while preserving essential spatiotemporal dynamics.

Keywords: Human-Object Interaction (HOI) recognition; learning from demonstration (LfD); imitation learning; temporal understanding; edge computing

Academic Editor: Tao Wang

Received: 10 October 2025
Revised: 26 October 2025
Accepted: 29 October 2025
Published: 31 October 2025

Citation: Pyaraka, J.C.; Isaksson, M.; McCormick, J.; Sutjipto, S.; Sukkar, F. Generalizable Interaction Recognition for Learning from Demonstration Using Wrist and Object Trajectories. *Electronics* **2025**, *14*, 4297. https://doi.org/10.3390/electronics14214297

1. Introduction

Learning from Demonstration (LfD) is gaining traction in robotics, enabling robots to acquire new skills by observing and imitating human instructors rather than relying on manual programming or hand-crafted control policies [1,2]. Unlike reinforcement learning, which often requires extensive trial-and-error interaction with the environment, LfD allows robots to leverage expert demonstrations as structured guidance, making it more practical for complex real-world manipulation tasks [3,4]. This approach is particularly appealing in domains where it is difficult to define explicit reward functions or symbolic task descriptions, such as manufacturing, service robotics, or assistive care. Over the past two decades, LfD has evolved from early trajectory-based imitation schemes to more advanced probabilistic, optimization-based, and deep learning frameworks, reflecting its growing importance in robotic autonomy [5].

The appeal of LfD lies not only in its intuitive teaching interface but also in its potential to lower the barriers for non-expert users to program collaborative robots. For example, LfD has been applied successfully to tasks such as handling compliant food materials, collaborative industrial tasks with robots, and programming contact-rich assembly operations [6]. In each of these settings, traditional programming or teleoperation approaches are often too brittle or cognitively demanding, whereas LfD provides a natural mechanism to encode human expertise directly into robotic policies. However, variability in demonstrations, diverse task contexts, and inconsistencies in human performance remain persistent obstacles to scalability [7]. These challenges have motivated a variety of recent approaches that attempt to refine, denoise, or generalize demonstrations, ranging from deep representation learning to meta-learning and one-shot generalization [8–10]. However, most of these methods remain data- and computation-intensive, limiting their practical use in real-time or resource-constrained robotics.

A core challenge in LfD is not only reproducing demonstrated trajectories but also interpreting the intent behind them, namely identifying what action or interaction is being demonstrated. Action recognition, and more specifically hand–object interaction (HOI) recognition, provides the semantic bridge between observed motion and task intent [11]. The accurate recognition of HOIs is essential for decomposing long demonstrations into reusable skills, enabling policy generalization, and transferring human demonstrations to robotic embodiments. Without robust interaction recognition, LfD risks degenerating into memorized trajectory mimicry, which limits adaptability to novel contexts. By contrast, models that can reliably identify interaction primitives such as reach, grasp, move, and release form the basis for hierarchical LfD pipelines, where high-level symbolic decisions are grounded in low-level control policies [12,13]. This has led to three main approaches for LfD representation:

- **Video-based approaches:**Video-based representations have become a prominent approach in LfD, leveraging the rich information contained in raw visual data to facilitate the acquisition of manipulation skills. The recent emergence of large-scale annotated video datasets, such as EPIC-KITCHENS [14] and Something-Something [15], has accelerated the development of models capable of extracting detailed spatiotemporal features from human demonstrations. Modern deep learning architectures including 3D Convolutional Neural Networks (3D CNNs), two-stream networks, and Transformer-based models have significantly enhanced the ability to learn from video by capturing both spatial and temporal dynamics [16,17]. Notably, action recognition frameworks such as Temporal Segment Networks (TSNs) [17] and Action Transformer Networks [16] explicitly model long-range temporal dependencies, which are crucial for segmenting and understanding extended manipulation tasks. Additionally, the EPIC-KITCHENS dataset [14] and Something-Something [15] have been central in benchmarking the generalization and robustness of video-based representation learning, promoting the design of systems that can recognize subtle variations in human–object interactions across diverse real-world settings. Recent robotics-focused advancements in this domain have tailored these techniques to the nuances of manipulation learning. For example, Yang et al. [18] proposed a two-stream framework tailored for robotic LfD, combining a grasp detection network to localize manipulable objects with a specialized video captioning network to interpret frame-level and local object dynamics. This method extends beyond generic video captioning by grounding visual semantics in object interactions relevant to robotic tasks. Similarly, Jia et al. [19] employed activity recognition and object detection within a vision-based pipeline, demonstrating the feasibility of learning multi-step skills from human videos using minimal supervision and domain knowledge transfer. Nevertheless, several challenges remain in deploying video-based LfD at scale. Achieving high accuracy often requires large volumes of annotated data and substantial computational resources, limiting adaptability to new tasks and real-time operation on embedded or resource-constrained robotic systems. Moreover, methods that rely solely on RGB or RGB-D data frequently face difficulties with occlusions, viewpoint variation, and background clutter, which are common in real-world environments [14,15,17]. Additionally, most video-based approaches lack explicit grounding in object pose or dynamic constraints, hindering effective policy transfer to physical robots and reducing generalization to scenarios outside the training distribution [18,19].

- **Pose-based approaches:** Pose-based methods form a fundamental strand of LfD research by representing human demonstrations primarily through the spatial and temporal evolution of skeletal joint positions. Advances in marker-based motion capture and, more recently, markerless pose estimation using deep learning methods such as OpenPose [20], or MediaPipe [21,22] have enabled the efficient extraction of full-body or hand skeletons from videos or sensor data. These representations enable the decomposition of demonstrations into interpretable sub-actions, such as reaching, grasping, or waving, thus facilitating modular skill learning and hierarchical policy design. Recent works have demonstrated the efficacy of skeleton-based recognition models including Graph Convolutional Networks (GCNs) [23], and Recurrent Neural Networks (RNNs) [24,25] in capturing fine-grained temporal and spatial features in human motion. For hand-centric manipulation, high-resolution fingertip and wrist trajectories support the nuanced analysis of actions relevant for robotic grasping and in-hand manipulation. Pose-based LfD has been successfully applied to collaborative robotics, teleoperation, and assistive scenarios, where gesture recognition and intent estimation are crucial [19]. Despite these strengths, pose-based methods exhibit several limitations for manipulation-centric LfD. Purely skeletal representations fail to capture object information, impeding the robust recognition of interactions where the object state or context is essential for task understanding. This can lead to ambiguities in action segmentation, especially in environments with clutter, occlusion, or similar gesture sequences involving different objects. Furthermore, the accuracy of pose estimation drops under challenging lighting, occlusion, or camera viewpoints, impacting learning in unstructured real-world settings. Consequently, there is growing interest in hybrid approaches that incorporate object localization and scene context alongside skeleton data to enrich demonstration encoding.
- **Object-centric approaches:** Object-centric approaches in LfD encode the state and dynamics of manipulated objects, typically as 6-DoF trajectories over time. By prioritizing object motion relative to task goals, these methods decouple learning from the specifics of human kinematics and instead emphasize the manipulation outcome. These representations are particularly effective in tasks such as assembly, insertion, and tool use, where success depends on changes in object states rather than human motion. Recent advances illustrate the potential of this paradigm. Sun et al. [26] proposed a pose-guided imitation learning framework for precise insertion tasks, representing demonstrations as relative Special Euclidean group SE(3) poses between source and target objects. With only 7–10 demonstrations, their method achieved millimeter-level precision, which was aided by a disentangled pose encoder and gated RGB-D fusion to handle noisy pose estimates. In parallel, Hsu et al. [27] introduced SPOT (SE(3) Pose Trajectory Diffusion), which is a trajectory diffusion model that synthesizes SE(3) object pose trajectories directly from demonstrations. By predicting

future object-centric trajectories rather than low-level actions, SPOT enabled closed-loop control and cross-embodiment generalization, transferring policies from human video to robotic execution.

While these categories often overlap—for example, pose-based methods may also rely on video input—they emphasize different sources of information and levels of abstraction when capturing human–object interactions. The object-centric LfD provides compact geometric encoding and robustness, and it often neglects intent and timing cues embedded in human wrist trajectories [28,29]. Recent advances in pose-object fusion for HOI detection have explored the keypoint-based interaction representations [30] and pose-aware feature refinements using full-body human keypoints [31]. However, these approaches are designed for static image detection rather than temporal sequence modeling, require extensive body pose estimation (17+ joints), and involve complex multi-stage processing, which are unsuitable for robotic learning. Motivated by these limitations, we propose a unified wrist–object representation that combines 2D wrist trajectories with full 6-DoF object poses into a single compact signal for interaction recognition in LfD. This representation captures both the demonstrator's intent and the resulting object state while avoiding the overhead of processing full video streams or dense skeletal inputs. By enabling efficient temporal modeling with reduced computational demands, the approach supports the robust recognition of high-level interactions and is empirically validated on GRasping Actions with Bodies (GRAB) [32] and First-Person Hand Action (FPHA) [33] datasets using Bidirectional Long Short-Term Memory (Bi-LSTM) with an attention and Transformer architectures. These architectures are used as representative temporal models to evaluate the proposed wrist–object representation rather than as new designs. The novelty of this work lies in the compact 10D wrist–object encoding, constructed from 2D wrist trajectories and 3D object poses, with object orientation expressed as a quaternion. This representation achieves high accuracy and real-time performance with minimal computation, effectively bridging pose-based and object-centric paradigms to provide a practical and scalable foundation for generalizable robot LfD.

The remainder of this paper is organized as follows: Section 2 presents the methodology, including the problem formulation for interaction recognition in LfD, the unified wrist–object feature extraction pipeline, and the temporal modeling architectures (Bi-LSTM with attention and Transformer). Section 3 describes the evaluation setup, detailing the GRAB [32] and FPHA [33] datasets, evaluation metrics encompassing classification accuracy, calibration measures, efficiency analysis, and the experimental design for comparing input representations and temporal models. Section 4 presents the results and discussion, analyzing the performance of different trajectory representations (5D, 6D, and 10D), comparing temporal architectures across both datasets, and evaluating computational efficiency and edge deployment feasibility. Finally, Section 5 concludes the paper with a summary of key findings and directions for future work.

2. Methodology

2.1. Problem Formulation

We formulate interaction recognition in LfD as a temporal sequence classification problem based on compact wrist and object trajectories. Each human demonstration is represented as a time-indexed sequence

$$D = \{(t, W(t), O(t)) \mid t = 1, \ldots, T\},$$

where t denotes the frame index, $W(t)$ is the 2D wrist coordinates extracted from the demonstrator, and $O(t) \in SE(3)$ represents the 6-DoF pose of the manipulated object.

The object pose is expressed as a translation vector $\mathbf{p} = (p_x, p_y, p_z)$ and a unit quaternion $\mathbf{q} = (q_x, q_y, q_z, q_w)$. The combined representation at time t can be expressed as

$$\mathbf{r}(t) = [\,t,\ \mathbf{w}(t),\ \mathbf{p}(t),\ \mathbf{q}(t)\,],$$

yielding a 10-dimensional feature vector per frame as shown in Figure 1. The objective is to learn a mapping

$$f : \{\mathbf{r}(t)\}_{t=1}^{T} \ \to\ \mathcal{Y},$$

where $\mathcal{Y}$ denotes the set of high-level interaction classes relevant to LfD (e.g., pick, pass, pour, open). The unified wrist–object representation establishes a direct correspondence between physical trajectory features and interaction semantics essential for LfD. The wrist trajectory $\mathbf{w}(t)$ encodes human motor intent patterns, for instance, the sharp approach-and-retreat characteristic of pick interactions versus the sustained positional engagement observed in inspect actions. These temporal motion signatures reflect deliberate behavioral strategies that distinguish interaction types at the level of human intent. Simultaneously, the object pose trajectory $O(t) = \mathbf{p}(t), \mathbf{q}(t)$ captures task-relevant affordance utilization, where translation patterns $\mathbf{p}(t)$ encode spatial manipulation outcomes while orientation dynamics $\mathbf{q}(t)$ reflect object-specific handling requirements. For example, drink actions exhibit distinctive upward translation coupled with tilting, while pick interactions show minimal orientation change during the initial lifting phases. This dual encoding ensures that the learned representation preserves both the demonstrator's intentional behavioral patterns and the resulting task-relevant object state changes, providing sufficient semantic information for robust robotic policy transfer and establishing a foundation for scalable and generalizable LfD. While this work focuses on recognizing high-level interactions, the same representation could be extended to decompose demonstrations into finer-grained atomic actions (e.g., reach, grasp, insert, align, fasten, clamp), which we leave as an avenue for further research.

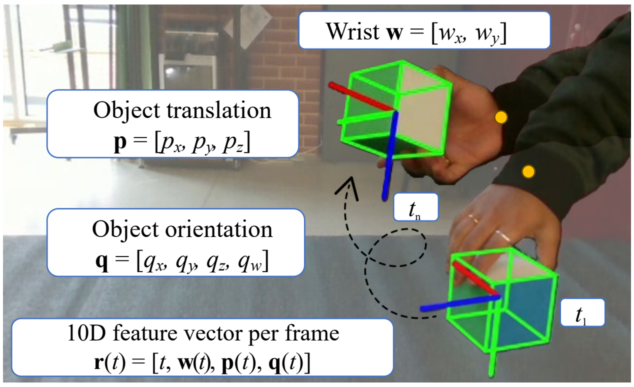
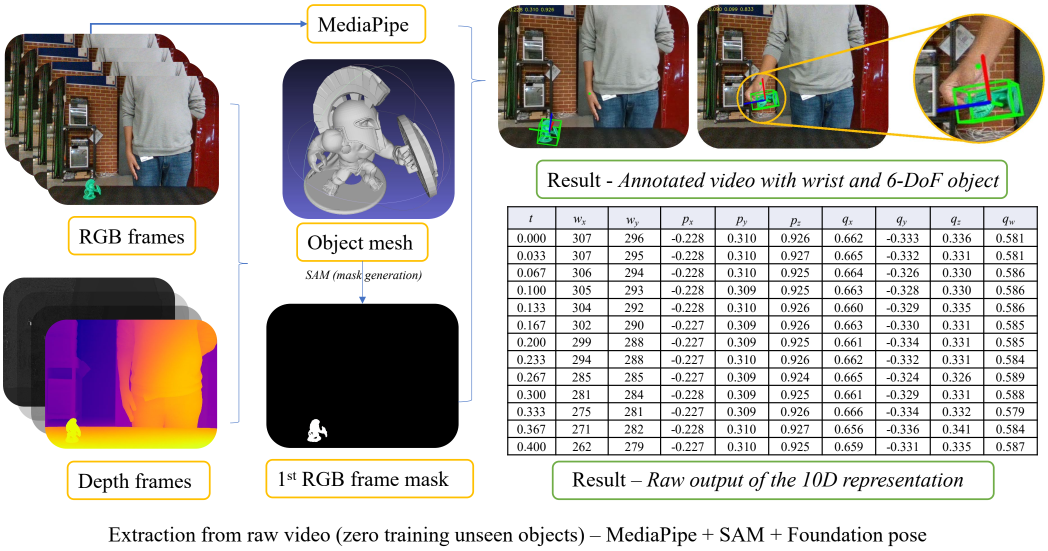

Wrist $\mathbf{w} = [w_x, w_y]$
Object translation $\mathbf{p} = [p_x, p_y, p_z]$
Object orientation $\mathbf{q} = [q_x, q_y, q_z, q_w]$
10D feature vector per frame $\mathbf{r}(t) = [t, \mathbf{w}(t), \mathbf{p}(t), \mathbf{q}(t)]$
t_n
t_1

Figure 1. Wrist and object trajectory representation used in this work. The wrist is represented by the 2D vector $\mathbf{w} = (w_x, w_y)$, while the object is described by its 3D translation $\mathbf{p} = (p_x, p_y, p_z)$ and orientation quaternion $\mathbf{q} = (q_x, q_y, q_z, q_w)$. Together with the time index t, these form the 10-dimensional feature vector $\mathbf{r}(t) = [t, \mathbf{w}(t), \mathbf{p}(t), \mathbf{q}(t)]$ used for interaction recognition.

2.2. Approach and Data Preprocessing

To obtain the unified wrist–object representation D described in Section 2.1 directly from raw video, we design a lightweight extraction pipeline. The complete process is illustrated in Figure 2. Object localization begins on the first RGB frame, where the user specifies a bounding region around the target object, which is converted into a binary mask. This mask, generated with the Segment Anything Model (SAM) [34], initializes FoundationPose [35], which subsequently maintains continuous 6-DoF object tracking across frames.


MediaPipe
RGB frames
Object mesh
SAM (mask generation)
Depth frames
1st RGB frame mask
Result - Annotated video with wrist and 6-DoF object
Result – Raw output of the 10D representation
Extraction from raw video (zero training unseen objects) – MediaPipe + SAM + Foundation pose

t	w_x	w_y	p_x	p_y	p_z	q_x	q_y	q_z	q_w
0.000	307	296	-0.228	0.310	0.926	0.662	-0.333	0.336	0.581
0.033	307	295	-0.228	0.310	0.927	0.665	-0.332	0.331	0.581
0.067	306	294	-0.228	0.310	0.925	0.664	-0.326	0.330	0.586
0.100	305	293	-0.228	0.309	0.925	0.663	-0.328	0.330	0.586
0.133	304	292	-0.228	0.310	0.926	0.660	-0.329	0.335	0.586
0.167	302	290	-0.227	0.309	0.926	0.663	-0.330	0.331	0.585
0.200	299	288	-0.227	0.309	0.925	0.661	-0.334	0.331	0.585
0.233	294	288	-0.227	0.310	0.926	0.662	-0.332	0.331	0.584
0.267	285	285	-0.227	0.309	0.924	0.665	-0.324	0.326	0.589
0.300	281	284	-0.228	0.309	0.925	0.661	-0.329	0.331	0.588
0.333	275	281	-0.227	0.309	0.926	0.666	-0.334	0.332	0.579
0.367	271	282	-0.228	0.310	0.927	0.656	-0.336	0.341	0.584
0.400	262	279	-0.227	0.310	0.925	0.659	-0.331	0.335	0.587

Figure 2. Unified wrist–object feature extraction pipeline for LfD. The process begins with synchronized RGB and depth frames from raw video. The wrist position is localized using MediaPipe [21], shown in green, while the object is segmented in the first RGB frame using SAM [34] to generate a binary mask. The segmented object mesh is then tracked across frames using FoundationPose [35], yielding the full 6-DoF object pose. On the top-right, overlaid detections from MediaPipe (wrist keypoints) and FoundationPose (object axes) are shown on selected video frames.

In parallel, wrist motion is tracked in real time using MediaPipe [21], providing 2D wrist coordinates for each frame. Occasional dropped detections are handled through interpolation or masking to ensure temporal continuity. FoundationPose [35] outputs the object's translation vector and orientation quaternion using RGB, depth, and camera intrinsics, providing a precise description of object state. The wrist and object signals are synchronized by the frame index and combined with the timestamp to form the compact trajectory representation $\mathbf{r}(t)$. Once extracted, sequences are standardized by downsampling frame rates and resampling or padding trajectories to a uniform length. Timestamps are converted to seconds, missing or infinite values are dropped, and noisy detections are interpolated. All features are scaled using MinMax normalization with the fitted scaler saved for identical transformation during inference and across validation/test splits to avoid leakage. A custom script fetches and merges per-action tables from MySQL schemas, automatically attaching action labels and producing uniform dataframes. Operating in real time at 30 FPS on a single GPU, this modular pipeline enables the consistent extraction of wrist and object trajectories from unconstrained video recordings, including unseen objects, and supports extension to new motion classes or sensing modalities.

2.3. Model Architectures and Training Strategy

This section presents the temporal modeling approaches applied to wrist and object trajectories using a custom Bi-LSTM with attention and a vanilla Transformer. Both architectures serve as representative sequence models for evaluating the proposed wrist–object representation rather than as direct competitors. Specifically, the Bi-LSTM was selected for its proven stability in processing bidirectional temporal sequences, effectively capturing both forward and backward dependencies in time-series data while maintaining robust performance across varying sequence lengths [36]. The Transformer was chosen as it represents one of the most recent and widely adopted paradigms for sequential reasoning, surpassing architectures such as GCNs and CNN-based encoders in capturing long-range temporal dependencies across diverse domains, including traffic prediction [37], human pose estimation [38], and spatiotemporal analysis [39]. Its self-attention mechanism directly addresses the limitations of recurrent models in modeling global sequence relationships, providing a complementary temporal modeling perspective to the Bi-LSTM. Together, these architectures enable a comprehensive evaluation of the proposed representation across both recurrent-based and attention-based modeling paradigms. However, architectural novelty is not the focus; instead, the study examines how effectively these temporal models can learn from the proposed 10D representation. This distinction emphasizes that the contribution lies in the representation design and its demonstrated efficiency and generalizability. To ensure fairness and reproducibility, both models follow identical preprocessing, normalization, and training protocols. Training is conducted using the Adam optimizer (learning rate = 0.001), a batch size of 32, and up to 50 epochs with early stopping (patience = 10) to prevent overfitting. Cross-entropy loss is used as the objective function across all experiments.

2.3.1. Bi-LSTM with Attention

The first sequential model is a bidirectional LSTM augmented with an attention mechanism. This configuration captures both forward and backward dependencies in the wrist and object trajectories, while attention allows the model to weight temporally salient frames, such as the onset of grasp or release. The architecture consists of two stacked Bi-LSTM layers, each with a hidden dimension of 100, which are followed by an attention layer that computes normalized relevance scores across the sequence. A weighted context vector is derived via softmax-based aggregation, normalized using layer normalization, regularized with dropout (0.5), and passed to a dense classification layer. The final dense layer outputs logits over the interaction label space.

2.3.2. Transformer

Most existing Transformer models for action or pose recognition either rely solely on full-body skeleton data, without incorporating object information or tracking, to model general human motions [40–45]. When addressing human–object interactions (HOIs), they typically process entire RGB or RGB-D frames to capture rich contextual cues [46]. Since few models explicitly leverage isolated hand and object trajectories (e.g., time-series keypoints or motion paths) for direct HOI recognition, we implement a simple vanilla Transformer classifier as a baseline for the evaluation of trajectory-based sequence modeling in resource-constrained setting. Input vectors are first projected into a hidden dimension of 48 via a linear layer, which is followed by two Transformer encoder layers with two attention heads each and a dropout of 0.3. Self-attention enables the model to capture both short-term temporal shifts and long-range dependencies without explicit recurrence. The resulting sequence embeddings are aggregated using mean pooling over non-padded tokens to produce a fixed-length summary vector that is classified by a final dense layer.

3. Evaluation Setup

3.1. GRasping Actions with Bodies (GRAB) Dataset

The GRAB dataset provides [32] a comprehensive collection of whole-body human grasping sequences, featuring detailed 3D shape and pose data for 10 subjects (5 male, 5 female) interacting with 51 everyday objects of varying geometries and sizes. Unlike conventional grasping datasets that focus exclusively on hand-object interactions, GRAB captures full 3D meshes of both human bodies and manipulated objects over time, along with computed body-object contact information, which is recorded in controlled laboratory conditions using high-precision motion capture systems. For our evaluation, we selected interaction classes based on the sample availability and relevance to LfD scenarios. We focused on five high-level interaction categories that contained at least 50 labeled sequences to ensure sufficient data for robust training and evaluation: pick, pass, inspect, offhand, and drink. This threshold-based selection strategy minimizes overfitting risks while providing an adequate representation of diverse manipulation contexts essential for robotic imitation learning applications. The distribution of training and test sequences is summarized in Table 1a.

Table 1. Train–test distribution of interaction categories in GRAB and FPHA datasets. (**a**) GRAB, (**b**) FPHA.

(**a**)

Dataset	**Split**	**Pick**	**Pass**	**Inspect**	**Offhand**	**Drink**
GRAB	Train	183	277	107	56	53
	Test	91	137	52	11	24

(**b**)

Dataset	**Split**	**Close**	**Open**	**Put**	**Pour**
FPHA	Train	60	61	26	61
	Test	15	14	7	14

3.2. First-Person Hand Action (FPHA) Dataset

The FPHA [33] dataset offers an egocentric perspective of hand–object interactions, containing over 105,000 RGB-D frames spanning 45 action categories and 26 different objects. However, only a specialized subset known as FPHA-HO contains the synchronized 6-DoF object pose annotations required for our unified wrist–object representation. This subset comprises approximately 21,501 frames focusing exclusively on four objects with consistent SE(3) pose tracking: juice bottle, liquid soap, milk bottle, and salt jar. From the FPHA-HO subset, we considered ten object–action combinations: open/close/pour actions with a juice bottle, liquid soap, and a milk bottle, plus put action with salt. For our high-level interaction recognition objective, these were abstracted into four semantic categories: open, close, pour, and put. This constraint to four objects was necessary due to the limited availability of synchronized wrist tracking and object SE(3) pose annotations required for our 10D input representation. The distribution of training and test sequences is summarized in Table 1b. Despite the smaller overall sample size compared to GRAB, this distribution provides adequate coverage of the four interaction categories for fair evaluation.

3.3. Dataset Variations

The combination of GRAB and FPHA provides complementary evaluation contexts for assessing our minimal wrist–object representation: GRAB offers controlled laboratory conditions with third-person viewpoints and whole-body context, while FPHA presents egocentric perspectives with different interaction semantics. These datasets were selected because they provide aligned RGB-D frames, object meshes and ground-truth 3D hand and object pose annotations, which were directly used to obtain wrist positions and 6-DoF object poses without additional estimation. While the datasets contain different interaction classes and cannot be used for cross-dataset generalization experiments, together they demonstrate the effectiveness of our unified representation across diverse viewpoints, interaction types, and recording modalities relevant to practical LfD applications.

3.4. Evaluation Metrics

To comprehensively assess the effectiveness of our wrist–object representation for interaction recognition, we adopt a combination of classification, calibration, efficiency, and statistical metrics across our experimental conditions.

- **Classification Accuracy and F1-Score:** The overall accuracy measures the proportion of correctly classified sequences, while the macro F1-score balances precision and recall across interaction classes. F1 is particularly important since some categories (e.g., offhandin GRAB, put in FPHA) contain fewer samples, and accuracy alone may overestimate performance on imbalanced datasets.
- **Probability Error Metrics:** In addition to categorical metrics (accuracy, F1), we report on the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) between predicted probability distributions and one-hot ground truth vectors. These measures quantify how far the predicted class probabilities deviate from the target labels, offering a finer-grained view of model calibration and prediction confidence. RMSE penalizes larger probability deviations more heavily, providing insight into model certainty and temporal consistency across different architectures and input representations.
- **Efficiency Metrics:** Since resource efficiency is essential for LfD deployment, we report on the training time (minutes), inference speed, and GPU/CPU memory usage (MB). These metrics demonstrate real-world deployability in embedded robotics settings where computational resources may be constrained.

3.5. Experiments

Our experimental evaluation consists of four complementary studies designed to validate the effectiveness of our minimal wrist–object representation.

- **Input Representation Comparison:** We evaluate different trajectory representations to assess the benefit of including object orientation information:
  - 5D: time + wrist coordinates (x, y) + object translation (x, y);
  - 6D: time + wrist coordinates (x, y) + object translation (x, y, z);
  - 10D: time + wrist coordinates (x, y) + full object pose $(x, y, z, q_x, q_y, q_z, q_w)$.

  This experiment tests whether the inclusion of orientations parameterized as quaternions significantly improves interaction recognition compared to translation-only object tracking. We include a 6D setting to isolate the contribution of object depth (z) before adding orientation. We report accuracy, F1, and probability calibration metrics (MAE, RMSE) for all representations across the GRAB [32] and FPHA [33] datasets.
- **Temporal Models:** We evaluate two temporal modeling architectures using both 10D inputs: Bi-LSTM with attention mechanism and Transformer encoder. This comparison assesses which temporal model better captures the dynamics of compact wrist–object sequences and provides more calibrated probability predictions for LfD applications. Both models are evaluated using identical training protocols and hyperparameters.
- **Efficiency Comparison:** Using the representation and model combination from Experiments 1–2, we conduct comprehensive efficiency analysis including training time, inference speed, memory footprint, and GPU vs. CPU performance. This experiment evaluates the computational requirements and deployability of our approach in resource-constrained robotics settings.

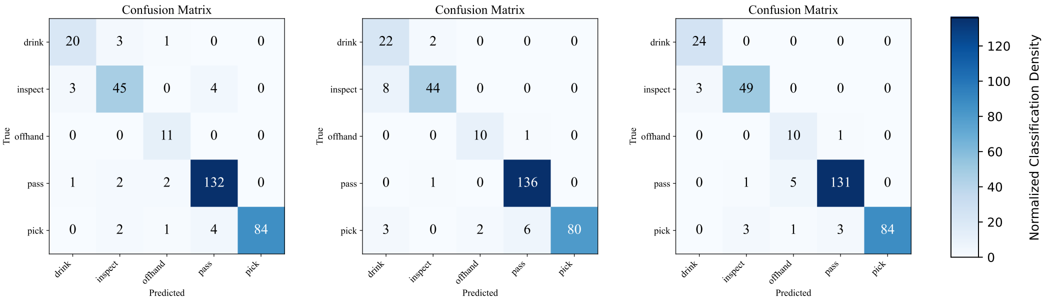
4. Results and Discussion

4.1. Input Representation Comparison

Table 2 presents the results of evaluating different trajectory representations on the GRAB [32] dataset using the Bi-LSTM with attention model (we restrict this experiment to GRAB, as it provides balanced classes and sufficient samples to make the 5D vs. 6D vs. 10D comparison meaningful). Figure 3 shows the corresponding confusion matrices. A clear progression is observed as additional object state information is included. The 5D baseline, which encodes only 2D wrist motion and planar object translation, already achieves strong performance with 92.7% accuracy and an F1-score of 92.8%. Introducing depth in the 6D representation further improves recognition, raising the accuracy to 93.1% and F1-score to 92.8%, confirming the importance of object distance cues for disambiguating interactions. Adding full $SE(3)$ orientation in the 10D representation yields the best performance, with 94.6% accuracy and 94.8% F1-score, while both MAE and RMSE are reduced, indicating more calibrated probability predictions. Although the 10D representation requires slightly longer training due to the additional orientation components, inference remains within real-time limits, and the overall memory footprint is minimal. This modest computational cost is justified by the significant gain in representational richness capturing both position and rotation, which are essential for accurate trajectory reproduction in LfD applications.

Table 2. Comparison of input representations on the GRAB dataset using Bi-LSTM with attention. The 10D wrist–object representation achieves the best performance across all metrics.

Input	Accuracy	F1-Score	MAE	RMSE	Training Time (s)	Inference Time (s)
5D	0.927	0.928	0.117	0.225	16.5	0.26
6D	0.931	0.928	0.107	0.204	71.0	0.65
10D	0.946	0.948	0.079	0.149	89.9	0.70



5D—Acc: 92.7% 6D—Acc: 92.7% 10D—Acc: 94.6%

Figure 3. Confusion matrices on the test set of the GRAB dataset for 5D, 6D, and 10D trajectory representations using the Bi-LSTM with attention model (see Table 1). Orientation information in the 10D representation reduces confusion between semantically similar interactions such as inspect and drink.

4.2. Temporal Models

Tables 3 and 4 summarize the performance of the Bi-LSTM with attention and Transformer architectures on both GRAB and FPHA datasets under GPU and CPU execution. Both models were trained using identical hyperparameters and evaluated with the 10D wrist–object representation to ensure fair comparison.

Table 3. Comparison of Bi-LSTM with attention and Transformer architectures on the GRAB dataset using 10D wrist–object inputs. Light gray rows denote GPU results.

Model	Device	Accuracy	F1-Score	MAE	RMSE	Training Time (s)	Inference Time (s)
Bi-LSTM (Attn.)	GPU	0.946	0.947	0.079	0.149	89.90	0.70
Bi-LSTM (Attn.)	CPU	0.905	0.909	0.142	0.307	612.88	0.68
Transformer	GPU	0.936	0.938	0.085	0.136	143.77	0.26
Transformer	CPU	0.918	0.923	0.130	0.257	734.40	0.27

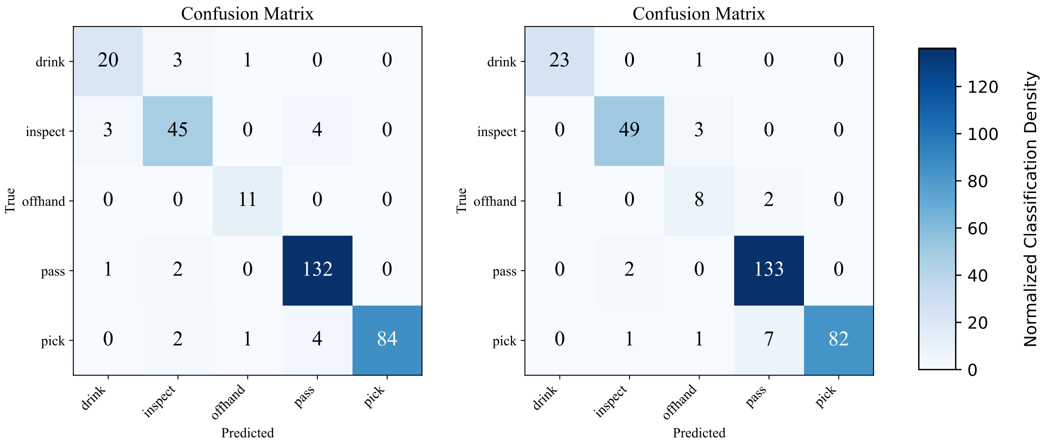
Table 4. Comparison of Bi-LSTM with attention and Transformer architectures on the FPHA dataset using 10D wrist–object inputs. Light gray rows denote GPU results.

Model	Device	Accuracy	F1-Score	MAE	RMSE	Training Time (s)	Inference Time (s)
Bi-LSTM (Attn.)	GPU	0.899	0.881	0.14	0.14	3.43	0.12
Bi-LSTM (Attn.)	CPU	0.919	0.900	0.09	0.09	62.35	0.05
Transformer	GPU	0.940	0.939	0.06	0.06	0.97	0.15
Transformer	CPU	0.961	0.959	0.04	0.04	14.07	0.03

4.2.1. GRAB Dataset

On the controlled GRAB [32] dataset, both architectures achieve high accuracy and F1-scores, confirming the effectiveness of the 10D wrist–object representation. The Bi-LSTM with attention attains the highest performance on GPU (0.946 accuracy, 0.947 F1), while the Transformer reaches a comparable 0.936 accuracy and 0.938 F1. The slight edge of the Bi-LSTM provides 1% improvement and indicates that sequential recurrence with attention weighting remains more effective for smooth, temporally consistent wrist–object trajectories captured in third-person, structured environments. Under CPU execution, the Transformer demonstrates stronger computational efficiency and stable accuracy (0.918 vs. 0.905 for Bi-LSTM) despite longer training time. However, Bi-LSTM exhibits lower calibration errors (MAE = 0.079 vs. 0.085 GPU) and reduced RMSE, indicating more confident probability estimates.

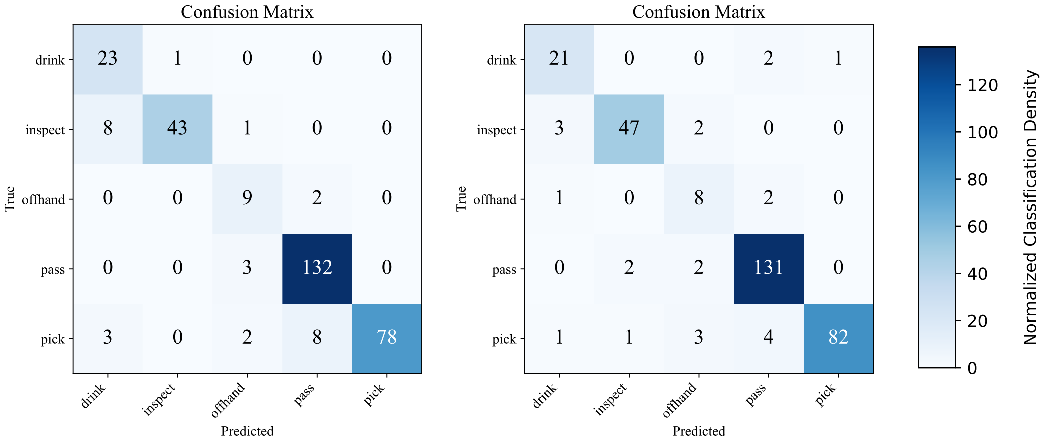
Overall, both models deliver inference times below 1 s, ensuring practical deployability in imitation learning contexts. Figure 4 shows the confusion matrices for both models under GPU and CPU execution on the GRAB test set. The Bi-LSTM with attention produces the most diagonally dominant matrices, confirming strong temporal consistency across all five interaction categories. Minor confusion arises primarily between inspect and drink, which share similar wrist trajectories but differ in object orientation and vertical motion. To further investigate this ambiguity, we analyze the learned temporal attention weights from the Bi-LSTM+Attention model. As shown in Figure 5, drink actions exhibit early attention peaks corresponding to rapid wrist–object convergence and lifting, while inspect actions show delayed attention aligned with sustained wrist positioning and gradual object rotation.



Bi-LSTM (GPU)—Acc: 93.3%

Transformer (GPU)—Acc: 94.2%

Figure 4. *Cont.*



Bi-LSTM (CPU)—Acc: 91.1% Transformer (CPU)—Acc: 92.3%

Figure 4. Confusion matrices on GRAB dataset: Bi-LSTM (GPU), Transformer (GPU), Bi-LSTM (CPU), and Transformer (CPU).

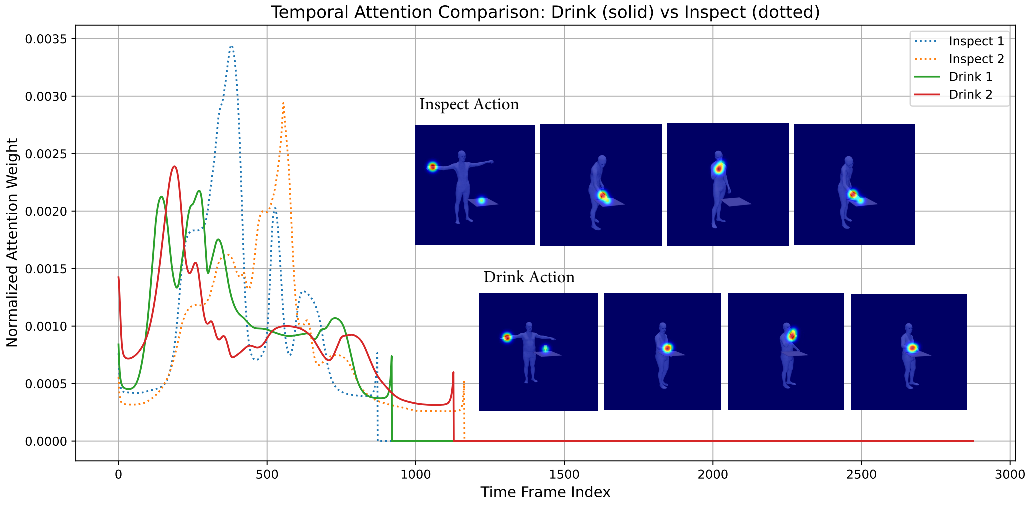


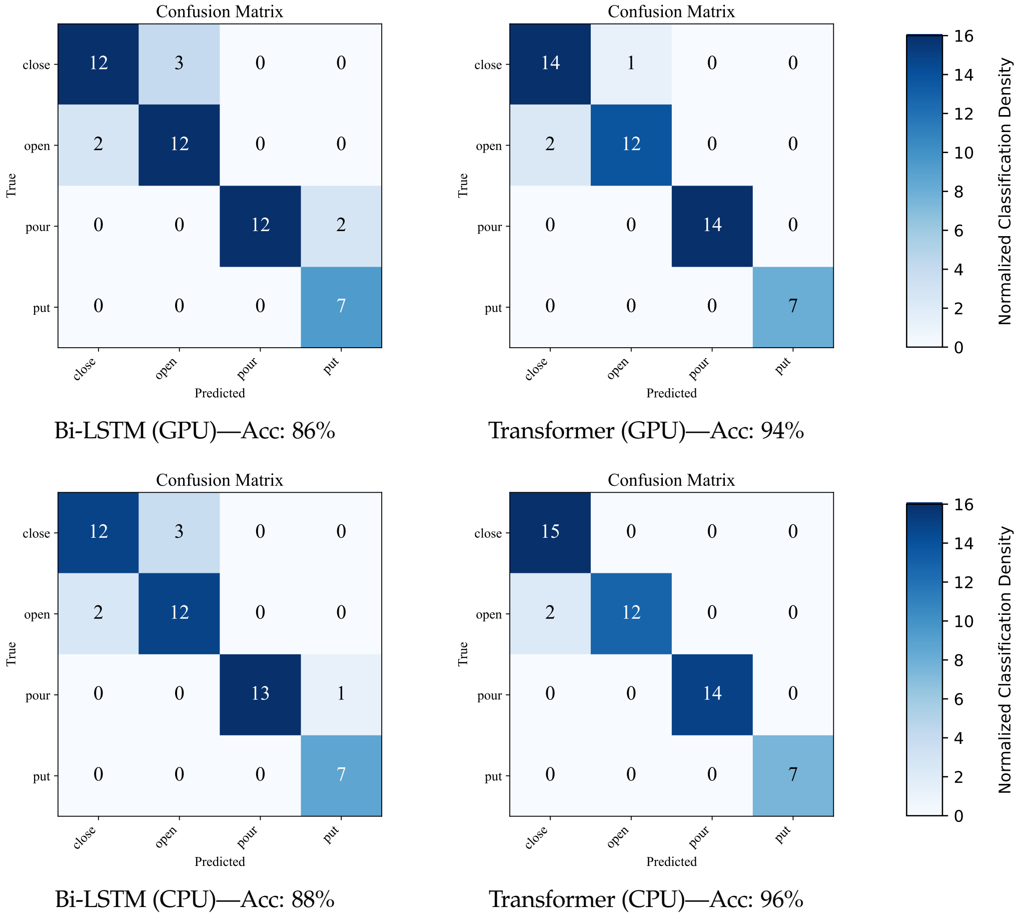
Figure 5. Temporal attention weights for two inspect (dotted, blue and orange) and two drink (solid, green and red) sequences from the GRAB test set, using the Bi-LSTM with attention model. The drink actions show early peaks aligned with wrist–object convergence and lifting, while the inspect actions exhibit delayed, sustained attention corresponding to object orientation changes. This demonstrates how the model learns to attend to task-relevant temporal segments in disambiguating similar interactions.

These patterns suggest that the model learns to associate specific temporal segments with semantic intent, validating the sufficiency of the 10D representation in capturing physical-to-semantic correspondences. Additionally, the slightly higher MAE and RMSE values observed for the Transformer model indicate less sharply calibrated probability distributions, aligning with the observed confusion between semantically close classes. Notably, CPU evaluation preserves the same structural trends as GPU inference, suggesting that the performance gap on CPU stems primarily from slower optimization dynamics rather than reduced discriminative capacity.

4.2.2. FPHA Dataset

The FPHA [33] dataset, being egocentric and noisier, poses greater variability in motion and occlusions. In this setting, the Transformer outperforms the Bi-LSTM, achieving 0.960 accuracy and 0.959 F1 on CPU with consistently low calibration errors (MAE = RMSE = 0.04). The Bi-LSTM remains competitive, scoring 0.919 on accuracy and attaining an F1-score of 0.90 on CPU but with a higher MAE (0.09). The Transformer shows a 4.2% advantage, suggesting that the Transformer's parallel self-attention mechanism generalizes better in unstructured, first-person settings where visual discontinuities can weaken the LSTM's sequential dependency modeling. Interestingly, both models show superior CPU performance on the smaller FPHA dataset, which is likely due to reduced overhead and more efficient memory utilization for the compact 10D representation when processing smaller batch sizes. These results highlight that incorporating orientation is the strongest contributor to recognition improvements, while adding depth alone provides only a modest refinement over the 5D baseline. The benefits are particularly relevant in distinguishing semantically close interactions. For example, inspect and drink may involve similar wrist trajectories, but the object's orientation trajectory differentiates holding-to-view from lifting-to-mouth. Although the 5D input already delivers strong recognition performance for high-level actions, enriching the representation with object pose further strengthens both accuracy and calibration. More importantly, it ensures that the extracted features retain sufficient task-relevant information for downstream LfD scenarios, where object translation and orientation are critical for faithfully reproducing human demonstrations on robotic platforms.

Figure 6 illustrates the confusion matrices for the FPHA dataset under the same evaluation settings. The Transformer demonstrates stronger separability between the open, close, and pour actions, maintaining sharper diagonal dominance and lower off-diagonal confusion, which is consistent with its lower MAE and RMSE values (0.04). In contrast, the Bi-LSTM occasionally confuses open and close, reflecting the temporal overlap between these gestures.



	close	open	pour	put
close	12	3	0	0
open	2	12	0	0
pour	0	0	12	2
put	0	0	0	7

Bi-LSTM (GPU)—Acc: 86%

	close	open	pour	put
close	14	1	0	0
open	2	12	0	0
pour	0	0	14	0
put	0	0	0	7

Transformer (GPU)—Acc: 94%

	close	open	pour	put
close	12	3	0	0
open	2	12	0	0
pour	0	0	13	1
put	0	0	0	7

Bi-LSTM (CPU)—Acc: 88%

	close	open	pour	put
close	15	0	0	0
open	2	12	0	0
pour	0	0	14	0
put	0	0	0	7

Transformer (CPU)—Acc: 96%

Figure 6. Confusion matrices on FPHA dataset: Bi-LSTM (GPU), Transformer (GPU), Bi-LSTM (CPU), and Transformer (CPU).

Despite this, both models achieve consistent recognition for put, suggesting that predominantly translational motions can be effectively captured without requiring complex orientation cues. Overall, the results confirm that the Transformer's self-attention mechanism generalizes better to subtle temporal variations in wrist–object coordination, whereas the Bi-LSTM remains slightly more sensitive to motion similarity among semantically close actions.

4.3. Efficiency Evaluation and Edge Feasibility

Table 5 summarizes the resource utilization and training efficiency of both architectures on GPU and CPU setups. All experiments were conducted on an RTX 4080 Laptop GPU (12 GB VRAM) with 24-thread CPU, demonstrating practical training within modest computational budgets.

- **Training Efficiency:** GPU training provides a clear speed advantage across both datasets with reductions of up to 7× compared to CPU training. For instance, Bi-LSTM training on the GRAB dataset completes in 89.9 s on GPU versus 612.9 s on CPU, while the Transformer requires 143.8 s and 734.4 s, respectively. The speedup arises from parallelized matrix operations and optimized batch execution on CUDA. Despite longer training times, CPU-only configurations remain practical for smaller datasets such as FPHA (62.3 s for Bi-LSTM), confirming that compact wrist–object representations do not demand large-scale GPU resources.
- **Memory Utilization:** GPU memory allocation during training remains well below 1.1 GB for both models with the total reserved memory between 2–4 GB. As shown in Table 6, memory requirements scale predictably with representation dimensionality (5D: 217 MB → 6D: 784 MB → 10D: 867 MB GPU allocation), yet all remain within practical deployment limits. This reflects PyTorch's (v 2.3.1) efficient dynamic allocation and suggests that both architectures can be trained even on mid-range GPUs (e.g., RTX 3060). CPU memory usage ranges from 1.3 GB to 1.9 GB, indicating that training and inference are feasible on standard desktop or embedded systems with $\geq$8 GB RAM. During inference, both models exhibit extremely lightweight requirements, reserving less than 300 MB of GPU memory and $\sim$1 GB of CPU memory across all representation sizes.
- **Inference Speed and Deployment:** Inference latency remains below one second on GRAB (0.26–0.70 s) and below 0.05 s on FPHA for both models, meeting real-time constraints for imitation learning. Transformer models exhibit slightly faster CPU inference (0.03 s vs. 0.05 s on FPHA), indicating their suitability for on-device or embedded deployment. In contrast, the Bi-LSTM shows marginally faster GPU training and superior probability calibration, making it advantageous for fine-tuning or incremental learning.
- **Edge Deployment Feasibility:** The observed efficiency profile confirms that our 10D wrist–object representation enables training and inference on both high-performance GPUs and compact CPUs without specialized hardware acceleration. As demonstrated in Table 7, with inference times and minimal memory footprint (<1.3GB), both Bi-LSTM and Transformer architectures can be deployed on lightweight robotic platforms or edge devices such as NVIDIA Jetson AGX Orin, Intel NUC, or ARM-based systems. The computational requirements are well within the capabilities of modern edge devices. For instance, the Jetson AGX Orin provides up to 275 TOPS of AI performance with 64 GB memory capacity, easily accommodating our method's resource demands. This efficiency makes the proposed framework practical for real-time LfD, where data-driven policy updates and online adaptation must occur within strict latency and resource limits.

Table 5. Resource consumption comparison across datasets and architectures.

Model	Dataset	Device	Training			Inference			Training Time (s)
			CPU Mem (MB)	GPU Alloc (MB)	GPU Rsrv (MB)	CPU Mem (MB)	GPU Alloc (MB)	GPU Rsrv (MB)	
Bi-LSTM (Attn.)	GRAB	GPU	1789.17	867.07	4400.00	1128.98	310.03	942.00	89.89
		CPU	1427.76	–	–	1068.07	–	–	612.88
	FPHA	GPU	1349.07	103.80	248.00	990.48	49.89	68.00	3.43
		CPU	811.54	–	–	577.94	–	–	62.35
Transformer	GRAB	GPU	1784.89	1086.02	2090.00	1192.53	136.88	222.00	143.77
		CPU	1924.19	–	–	1242.82	–	–	734.40
	FPHA	GPU	1381.73	72.75	114.00	1073.90	12.87	26.00	0.97
		CPU	822.51	–	–	575.06	–	–	14.07

Table 6. Memory consumption comparison across input representations (GRAB dataset, Bi-LSTM).

Representation	Training			Inference		
	CPU (MB)	GPU Alloc (MB)	GPU Rsrv (MB)	CPU (MB)	GPU Alloc (MB)	GPU Rsrv (MB)
5D	1410.15	217.35	860.00	1016.49	93.02	238.00
6D	1612.19	783.81	3030.00	1072.69	207.22	744.00
10D	1789.17	867.07	4400.00	1128.98	310.03	942.00

Table 7. Deployment efficiency summary: Inference time and peak memory usage for the models across datasets and devices. Inference times reflect processing of the entire test set per model–dataset combination.

Dataset	Model	Device	Inference Time (s)	Peak Memory (MB)
GRAB	Bi-LSTM	GPU	0.70	CPU: 1128.98/GPU: 310.03
		CPU	0.68	CPU: 1068
	Transformer	GPU	0.26	CPU: 1192.53/GPU: 136.88
		CPU	0.27	CPU: 1242.82
FPHA	Bi-LSTM	GPU	0.12	CPU: 990.48/GPU: 49.89
		CPU	0.05	CPU: 577.94
	Transformer	GPU	0.15	CPU: 1073.90/GPU: 12.87
		CPU	0.03	CPU: 575.06

5. Conclusions

This study presents a generalizable representation for LfD that encodes human–object interactions using only the wrist trajectory and the object's full SE(3) pose. A unified feature-extraction pipeline was developed to derive these trajectories directly from RGB-D videos by integrating MediaPipe-based wrist tracking with FoundationPose based 6-DoF object estimation. Evaluations on the GRAB [32] and FPHA [33] datasets demonstrate that this compact 10D representation effectively captures both human intent and task outcome while maintaining high recognition accuracy and calibration. Consistent results across Bi-LSTM and Transformer architectures confirm the framework's robustness and highlight its balance between interpretability and computational efficiency. By unifying pose-based and object-centric perspectives, the approach allows wrist trajectories and object poses to jointly encode the spatiotemporal dynamics of interactions in a minimal yet expressive form. The framework achieves practical deployability with less than a second inference times, minimal memory footprint (<1 GB), and successful operation on both GPU and CPU platforms, enabling implementation on edge devices such as NVIDIA Jetson or Intel NUC systems. This efficiency addresses key constraints in practical robotic deployment while avoiding the computational overhead of full video processing as feature inputs.

Future research will extend this representation toward fine-grained, atomic action recognition, decomposing demonstrations into reach, grasp, align, and release phases based on wrist–object trajectories. In parallel, the trajectory-level understanding will be integrated into closed-loop robotic manipulation, enabling the direct replication of demonstrated actions on physical manipulators. To facilitate broader progress in imitation learning within manufacturing, a dedicated multi-view dataset is currently being developed to capture tool-based industrial actions with synchronized wrist and object trajectories. This dataset will provide the empirical foundation for advancing fine-grained action segmentation and transfer learning in robotic manipulation. Together, these developments will advance the framework from interaction recognition toward full end-to-end LfD, where robots learn both what to do and how to execute it from minimal human data. Additionally, we are exploring extensions of the 10D representation to incorporate dynamic features such as velocity, angular velocity, and contact-aware cues while also examining their impact on training and deployment efficiency for future integration into physically grounded LfD pipelines.

Author Contributions: Conceptualization, J.C.P., M.I., J.M., S.S. and F.S.; Methodology, J.C.P., M.I. and J.M.; Software, J.C.P.; Validation, J.C.P.; Formal analysis, J.C.P.; Investigation, J.C.P.; Data curation, J.C.P.; Writing—original draft, J.C.P.; Writing—review & editing, M.I., J.M. and S.S.; Supervision, M.I., J.M. and S.S.; Project administration, M.I.; Funding acquisition, M.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Australian Government's ARC Industrial Transformation Training Centre (ITTC) for Collaborative Robotics in Advanced Manufacturing under Grant IC200100001.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Schaal, S. Learning from Demonstration. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 1996 ; Volume 9.
2. Argall, B.D.; Chernova, S.; Veloso, M.; Browning, B. A survey of robot learning from demonstration. *Robot. Auton. Syst.* **2009**, *57*, 469–483. [CrossRef]
3. Billard, A.; Calinon, S.; Dillmann, R.; Schaal, S. Robot Programming by Demonstration. In *Springer Handbook of Robotics*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 1371–1394. [CrossRef]
4. Chernova, S.; Thomaz, A.L. Modes of Interaction with a Teacher. In *Robot Learning from Human Teachers*; Chernova, S.; Thomaz, A.L., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2014; pp. 17–24. [CrossRef]
5. Ravichandar, H.; Polydoros, A.S.; Chernova, S.; Billard, A. Recent Advances in Robot Learning from Demonstration. *Annu. Rev. Control. Robot. Auton. Syst.* **2020**, *3*, 297–330. [CrossRef]
6. Osa, T.; Pajarinen, J.; Neumann, G.; Bagnell, J.A.; Abbeel, P.; Peters, J. An Algorithmic Perspective on Imitation Learning. *Found. Trends Robot.* **2018**, *7*, 1–179. [CrossRef]
7. Cederborg, T.; Li, M.; Baranes, A.; Oudeyer, P.Y. Incremental local online Gaussian Mixture Regression for imitation learning of multiple tasks. In Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan, 18–22 October 2010 ; pp. 267–274. [CrossRef]
8. Rahmatizadeh, R.; Abolghasemi, P.; Bölöni, L.; Levine, S. Vision-Based Multi-Task Manipulation for Inexpensive Robots Using End-To-End Learning from Demonstration. *arXiv* **2018**, arXiv:1707.02920. [CrossRef]
9. Finn, C.; Abbeel, P.; Levine, S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *arXiv* **2017**, arXiv:1703.03400. [CrossRef]
10. Duan, Y.; Andrychowicz, M.; Stadie, B.; Jonathan Ho, O.; Schneider, J.; Sutskever, I.; Abbeel, P.; Zaremba, W. One-Shot Imitation Learning. In *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Long Beach, CA, USA, 2017; Volume 30.
11. Konidaris, G.; Barto, A. Skill Discovery in Continuous Reinforcement Learning Domains using Skill Chaining. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Vancouver, BC, Canada, 2009; Volume 22.
12. Shan, D.; Geng, J.; Shu, M.; Fouhey, D.F. Understanding Human Hands in Contact at Internet Scale. *arXiv* **2020**, arXiv:2006.06669. [CrossRef]
13. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. SlowFast Networks for Video Recognition. *arXiv* **2019**, arXiv:1812.03982. [CrossRef]

14. Damen, D.; Doughty, H.; Farinella, G.M.; Furnari, A.; Kazakos, E.; Ma, J.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; et al. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. *Int. J. Comput. Vis.* **2022**, *130*, 33–55. [CrossRef]
15. Goyal, R.; Kahou, S.E.; Michalski, V.; Materzyńska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. The "something something" video database for learning and evaluating visual common sense. *Proc. IEEE Int. Conf. Comput. Vis.* **2017**, arXiv:1706.04261. [CrossRef]
16. Girdhar, R.; João Carreira, J.; Doersch, C.; Zisserman, A. Video Action Transformer Network. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019. [CrossRef]
17. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal Segment Networks for Action Recognition in Videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2740–2755. [CrossRef]
18. Yang, S.; Zhang, W.; Lu, W.; Wang, H.; Li, Y. Learning Actions from Human Demonstration Video for Robotic Manipulation. *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* **2019**, 1805–1811. [CrossRef]
19. Jia, Z.; Lin, M.; Chen, Z.; Jian, S. Vision-based Robot Manipulation Learning via Human Demonstrations. *arXiv* **2020**, arXiv:2003.00385. [CrossRef]
20. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.E.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**,*43*, 172–186. [CrossRef]
21. Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.L.; Yong, M.G.; Lee, J.; et al. MediaPipe: A Framework for Building Perception Pipelines. *arXiv* **2019**, arXiv:1906.08172. [CrossRef]
22. Singh, A.K.; Kumbhare, V.A.; Arthi, K. Real-Time Human Pose Detection and Recognition Using MediaPipe. In *Soft Computing and Signal Processing*; Reddy, V.S., Prasad, V.K., Wang, J., Reddy, K., Eds.; Springer Nature: Singapore, 2022; pp. 145–154. [CrossRef]
23. Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. *Proc. AAAI Conf. Artif. Intell.* **2018**, *32*, 1–8. [CrossRef]
24. Belluzzo, B.; Marana, A.N. Human Action Recognition Based on 2D Poses and Skeleton Joints. In *Intelligent Systems*; Xavier-Junior, J.C., Rios, R.A., Eds.; Springer International Publishing: Springer, Cham, 2022; pp. 71–83. [CrossRef]
25. Song, S.; Lan, C.; Xing, J.; Zeng, W.; Liu, J. Spatio-Temporal Attention-Based LSTM Networks for 3D Action Recognition and Detection. *IEEE Trans. Image Process.* **2018**, *27*, 3459–3471. [CrossRef] [PubMed]
26. Sun, H.; Wang, Y.; Zhou, Z.; Wang, S.; Yang, H.; Sun, J.; Cao, Q. Exploring Pose-Guided Imitation Learning for Robotic Precise Insertion. *arXiv* **2025**, arXiv.2505.09424. [CrossRef]
27. Hsu, C.C.; Wen, B.; Xu, J.; Narang, Y.; Wang, X.; Zhu, Y.; Biswas, J.; Birchfield, S. SPOT: SE(3) Pose Trajectory Diffusion for Object-Centric Manipulation. In Proceedings of the 2025 IEEE International Conference on Robotics and Automation (ICRA), Atlanta, GA, USA, 19–23 May 2025; pp. 4853–4860. [CrossRef]
28. Dang, H.; Allen, P.K. Robot learning of everyday object manipulations via human demonstration. In Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan, 18–22 October 2010; pp. 1284–1289. [CrossRef]
29. Carfì, A.; Patten, T.; Kuang, Y.; Hammoud, A.; Alameh, M.; Maiettini, E.; Weinberg, A.I.; Faria, D.; Mastrogiovanni, F.; Alenyà, G.; et al. Hand-Object Interaction: From Human Demonstrations to Robot Manipulation. *Front. Robot. AI* **2021**, *8*, 714023. [CrossRef]
30. Wang, T.; Yang, T.; Danelljan, M.; Khan, F.S.; Zhang, X.; Sun, J. Learning Human-Object Interaction Detection Using Interaction Points. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020.
31. Wu, E.Z.Y.; Li, Y.; Wang, Y.; Wang, S. Exploring Pose-Aware Human-Object Interaction via Hybrid Learning. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 17815–17825. [CrossRef]
32. Taheri, O.; Ghorbani, N.; Black, M.J.; Tzionas, D. GRAB: A Dataset of Whole-Body Human Grasping of Objects. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020.
33. Garcia-Hernando, G.; Yuan, S.; Baek, S.; Kim, T.K. First-Person Hand Action Benchmark with RGB-D Videos and 3D Hand Pose Annotations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 409–419.
34. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 4015–4026.
35. Wen, B.; Yang, W.; Kautz, J.; Birchfield, S. FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–18 June 2024; pp. 18677–18687.

36. Siami-Namini, S.; Tavakoli, N.; Namin, A.S. The Performance of LSTM and BiLSTM in Forecasting Time Series. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December 2019; pp. 3285–3292. [CrossRef]
37. Li, A.; Xu, Z.; Li, W.; Chen, Y.; Pan, Y. Urban Signalized Intersection Traffic State Prediction: A Spatial-Temporal Graph Model Integrating the Cell Transmission Model and Transformer. *Appl. Sci.* **2025**, *15*, 2377. [CrossRef]
38. Mehraban, S.; Adeli, V.; Taati, B. MotionAGFormer: Enhancing 3D Human Pose Estimation With a Transformer-GCNFormer Network. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2024; pp. 6920–6930.
39. Zhang, J.; Yang, Y.; Wu, X.; Li, S. Spatio-temporal transformer and graph convolutional networks based traffic flow prediction. *Sci. Rep.* **2025**, *15*, 24299. [CrossRef]
40. Xin, W.; Liu, R.; Liu, Y.; Chen, Y.; Yu, W.; Miao, Q. Transformer for Skeleton-based action recognition: A review of recent advances. *Neurocomputing* **2023**, *537*, 164–186. [CrossRef]
41. Do, J.; Kim, M. SkateFormer: Skeletal-Temporal Transformer for Human Action Recognition. In *Computer Vision—ECCV 2024*; Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G., Eds.; Lecture Notes in Computer Science; Springer Nature: Cham, Switzerland, 2025; Volume 15099, pp. 401–420. [CrossRef]
42. Chen, D.; Chen, M.; Wu, P.; Wu, M.; Zhang, T.; Li, C. Two-stream spatio-temporal GCN-transformer networks for skeleton-based action recognition. *Sci. Rep.* **2025**, *15*, 4982. [CrossRef]
43. Pang, Y.; Ke, Q.; Rahmani, H.; Bailey, J.; Liu, J. IGFormer: Interaction Graph Transformer for Skeleton-Based Human Interaction Recognition. In *Computer Vision—ECCV 2022*; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Lecture Notes in Computer Science; Springer Nature: Cham, Switzerland, 2022; Volume 13685, pp. 605–622. [CrossRef]
44. Hu, Q.; Liu, H. Multi-Modal Enhancement Transformer Network for Skeleton-Based Human Interaction Recognition. *Biomimetics* **2024**, *9*, 123. [CrossRef]
45. Wang, X.; Jiang, X.; Zhao, Z.; Wang, K.; Yang, Y. Exploring interaction: Inner-outer spatial–temporal transformer for skeleton-based mutual action recognition. *Neurocomputing* **2025**, *636*, 130007. [CrossRef]
46. Wang, T.; Anwer, R.M.; Khan, M.H.; Khan, F.S.; Pang, Y.; Shao, L.; Laaksonen, J. Deep Contextual Attention for Human-Object Interaction Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 5693–5701. [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.