2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

DexSkills: Skill Segmentation Using Haptic Data for Learning Autonomous Long-Horizon Robotic Manipulation Tasks

Xiaofeng Mao[1,†] and Gabriele Giudici[2,†],
Claudio Coppola[3], Kaspar Althoefer[2], Ildar Farkhatdinov[2,4], Zhibin Li[5], Lorenzo Jamone[2]

***Abstract*— Effective execution of long-horizon tasks with dexterous robotic hands remains a significant challenge in real-world problems. While learning from human demonstrations has shown encouraging results, they require extensive data collection for training. Hence, decomposing long-horizon tasks into reusable primitive skills is a more efficient approach. To achieve so, we developed DexSkills, a novel supervised learning framework that addresses long-horizon dexterous manipulation tasks using primitive skills. DexSkills is trained to recognize and replicate a select set of skills using human demonstration data, which can then segment a demonstrated long-horizon dexterous manipulation task into a sequence of primitive skills to achieve one-shot execution by the robot directly. Significantly, DexSkills operates solely on proprioceptive and tactile data, i.e., haptic data. Our real-world robotic experiments show that DexSkills can accurately segment skills, thereby enabling autonomous robot execution of a diverse range of tasks.**

I. INTRODUCTION

Humans show remarkable dexterity and adaptability in performing manipulation tasks across various environments, this is attributed to the inherent capabilities of the human hand. However, enabling robotic dexterous manipulation with multi-fingered hands remains challenging due to the high-dimensional action space and occlusion during manipulation. Recent research efforts have focused on addressing this challenge through the model-based control methods [1], [2], [3], as well as learning-based approaches, such as learning from human demonstrations [4] or training via model-free reinforcement learning directly from scratch [5], [6], [7], [8]. However, these methods rely on complex mathematical models, require customized rewards designed by humans, demand extensive training time, and necessitate substantial amounts of demonstration data [9]. Consequently, the training processes become inefficient and labor-intensive, hindering robots from achieving dexterous manipulation.

Multi-modal information is crucial in enabling robot dexterous manipulation. In routine tasks like pick-and-place operations, visual cues provide information on the position and shape of the object, aiding the robot during its execution. However, the availability of accurate visual information may be compromised due to factors such as poor lighting conditions or occlusions. In addition, certain aspects of objects and tasks, especially in contact-rich manipulation tasks, are better encoded by tactile and force sensing [10], [11]. Tactile feedback offers valuable insights into the contact dynamics between the robotic fingers and the manipulated object, enhancing the potential for accurate and stable manipulation. Hence, numerous studies are currently investigating the viability of leveraging tactile information as an alternative or complementary source [12], [13], [14], [15], [16].

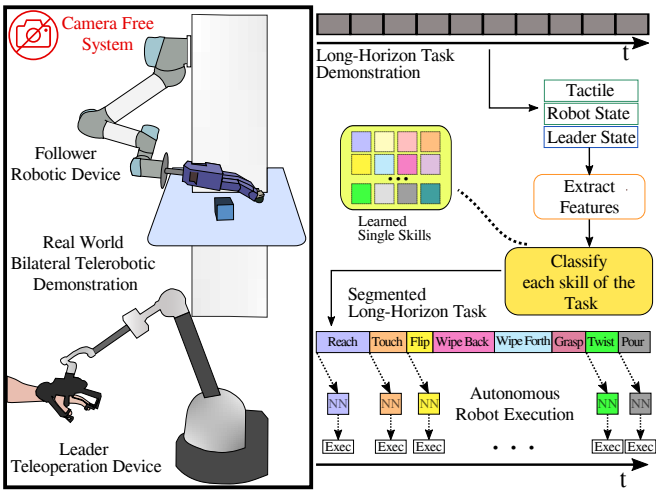


Fig. 1: Overview of the proposed long-horizon task segmentation approach. Individual skills are segmented and classified at each temporal window of the demonstration. The demonstrations are collected via the teleoperation system (left) developed in [17].

Learning from human demonstrations makes it possible to learn from real-world human demonstrations, avoiding the need for intricate dynamic models, well defined reward functions, and reliable simulated environments [18]. On the other hand, these approaches often demand a considerable amount of data and human effort [19]. Moreover, for long-horizon tasks, these methods are prone to failure due to the accumulation of compounding errors and necessitate model retraining for each new demonstrated task, even if many task phases are in common between tasks [20], [21].

Therefore, we propose DexSkills, a system that learns primitive skills from human demonstrations, and uses such

Xiaofeng Mao and Gabriele Giudici contributed equally to this work and share first authorship.

[1]University of Edinburgh, email: xiaofeng.mao@ed.ac.uk
[2]ARQ (the Centre for Advanced Robotics @ Queen Mary), School of Engineering and Materials Science, Queen Mary University of London, London, E14NS, UK (emails: {g.giudici, k.althoefer, i.farkhatdinov, l.jamone}@qmul.ac.uk).
[3]Humanoid AI, email: ccop@thehumanoid.ai
[4]School of Biomedical Engineering and Imaging Sciences, King's College London, UK.
[5]University College London, email: alex.li@ucl.ac.uk

The study is funded by the UKRI EPSRC grants EP/R02572X/1 (NCNR), EP/V035304/1 (q-Arena) and Queen Mary University of London Ph.D. scholarship to G. Giudici.

knowledge to segment a human-demonstrated long-horizon task in a sequence of skills that the robot can execute autonomously. This requires less data collection efforts with respect to learning each possible long-horizon task from human demonstration. Notably, DexSkills relies only on proprioceptive and tactile data (i.e. haptic data). In particular, the contributions of this work are:

1) We introduce a novel supervised representation learning framework where the latent features from haptic data are jointly trained via an auto-regressive autoencoder and a label decoder. Our qualitative result proves that by capturing the latent dynamics of primitive skills, the performance of segmenting unseen long-horizon tasks into skill sequences is significantly improved;
2) We propose a set of 20 core primitive skills specifically designed to address robot dexterous manipulation tasks. We trained the autonomous robot controller for each skill using Multi-Layer Perceptron (MLP) and demonstrated that by recombining these trained core primitive skills, the robot is able to compose and achieve long-horizon dexterous tasks effectively and successfully;
3) To facilitate future research, we open-sourced our code[1] and provided a labeled dataset comprising 20 primitive skills and 20 long-horizon tasks, which were obtained from human teleoperated demonstrations.

DexSkills has developed a primary set of reusable primitive skills trained from haptic human demonstrations, which can be reused and re-combined to represent various compositions of long-horizon tasks. DexSkills achieves remarkable segmentation accuracy of 91% for unseen tasks and enables autonomous robot execution of diverse tasks using only proprioceptive and tactile data.

II. Related Works

Imitation learning based robot dexterous manipulation: Robot dexterous manipulation has recently gained increasing attention [22]. Imitation learning (IL) is a well-established method that accelerates the skill learning process in robots by leveraging human experience [23]. In [24], human demonstrations are captured within a virtual environment, employing learning-based methods for feedback control in executing non-prehensile manipulations The research [25] proposes to train an inverse dynamics model to predict actions for state-only demonstration [25]. In [26], an RGB camera is used to observe a human operator teleoperate a robot and train the policy with the imitation learning method.

Long-horizon imitation learning: For long-horizon manipulation tasks, traditional IL methods might not be sufficient and could fail due to the accumulation of compounding errors. Breaking down the long-horizon manipulation task into multiple sub-tasks is a widely adopted strategy for addressing this challenge. By decomposing the IL for manipulation task into visual servoing and behaviour replay, [27] has effectively facilitated one-shot imitation learning for real-world daily tasks. Additionally, the strategy of automatic waypoint extraction to segment the entire trajectory significantly reduces the BC horizon, effectively mitigating the problem of error accumulation [21]. Hierarchical imitation learning serves as a prominent strategy to solve long-horizon manipulation tasks. By training a high-level policy for skill selection and a low-level policy for precise motor control, robots are capable of handling a wide range of dexterous manipulation tasks, including activities like cable routing [20], drink pouring [28] and bi-manual cooperation task [29], [30].

Dexterous manipulation with tactile sensing: Recently, some works have investigated using tactile sensing to mitigate occlusion during manipulation tasks with robot dexterous hand. By overlaying the binary force sensor to the one side of the whole robot hand, the work [13] has successfully facilitated the object in-hand rotation task. With the same robot hand setup, the study [14] demonstrated the capability of employing tactile sensors to search for and locate target objects, subsequently manipulating these objects to perform daily tasks. Similarly, [31] use the tactile sensor Uskin [32] and proposes to pre-train a tactile encoder to extract features from high-dimensional tactile information, which enables the robot to perform tasks such as book opening, bowl and cup unstacking. The work [15] investigates using sparse tactile feedback to localize, identify and grasp novel objects without any visual feedback.

Primitive Skills: Daily life manipulation tasks often comprise a variety of primitive skills. Identifying the sequence in which these skills combine for any long-horizon task is crucial for facilitating their reuse, thereby broadening the scope of tasks that can be accomplished. Retrieval skills from long sequence demonstration can be achieved by using supervised [33], [34] and unsupervised method [35], [36], [37], [38], [39]. For the reuse of the primitive skills, [40] proposes to train a universal agent with 12 unique skills that are capable of multi-task manipulation by using semantic augmentation and action representation. The research in [41] proposes a hybrid hierarchical learning framework named ROMAN, which combines behavioural cloning (BC), reinforcement learning (RL), and IL – where a central manipulation network learns how to produce the appropriate sequential actions for various sub-skills networks, orchestrating multiple skills to solve complex long-horizon manipulation tasks. The work [42] proposes to chain multiple dexterous policies for achieving long-horizon dexterous manipulation tasks by defining a feasibility function. Different from previous works, we introduce primitive skills specifically designed for dexterous manipulation tasks, aimed at achieving unseen long-horizon tasks by predicting the sequence of primitive skills combination. Inspired by the work [33], we introduce a specialized set of features designed for telerobot dexterous manipulation, aiming to substantially enhance classifier performance through the integration of auto-regressive autoencoder (AE). Moreover, we have developed and trained a distinct skill policy for each primitive, employing the predicted skill sequence to successfully execute long-horizon manipulation tasks.

[1] https://github.com/ARQ-CRISP/DexSkills

III. METHODS

In this study, our objective is to explore methods for accomplishing imitation learning for long-horizon dexterous manipulation tasks. The framework for learning long-horizon dexterous manipulation tasks is illustrated in Fig 1. We start by proposing primitive skills for tasks involving the use of a robotic arm and dexterous hand. Our research focuses on learning these primitive skills with haptic data, segmenting the skill sequence from unseen long-horizon dexterous manipulation tasks, and executing the task by sequentially performing the identified skill segments.

A. Learning Features

For the robotic setup, which includes a robotic arm, a dexterous hand, and tactile sensors, we propose a set of features designed to differentiate primitive skills. These features are summarized as follows:

End Effector (EE) Information: For the features related to the EE, we utilize both position and velocity information to accurately capture the movements. Additionally, to provide a clearer representation of the movement of EE direction, we incorporate the EE direction as a feature. This direction is defined with a value into the set $[-1, 0, 1]$. Specifically, when the velocity exceeds $0.02m/s$, the EE direction is designated as 1 to indicate forward movement. Conversely, if the velocity is less than $-0.02m/s$, the direction is set to -1, indicating backward movement. In scenarios where the velocity of the EE falls within the threshold, the direction is set to 0, denoting a stationary state. These features yield a detailed understanding of the movement of the EE throughout the execution of tasks.

- EE Pose: $[x, y, z, \alpha, \beta, \gamma]$
- EE Velocity: $[\dot{x}, \dot{y}, \dot{z}, \dot{\alpha}, \dot{\beta}, \dot{\gamma}]$
- EE Direction: $[-1, 0, 1]$

Allegro Hand (AH) Information: The features of the Allegro Hand (AH) include the AH joint state, fingertip positions, velocities, as well as position and velocity covariance. To effectively capture the movement correlations between each finger, the *Log* of the upper triangle of the covariance matrix is utilized. This approach is chosen for its efficiency in representing finger movements during manipulation tasks.

- AH Joint State: $[16 \times 1]$
- Fingerfitp Position: $[x, y, z]$
- Fingertips Velocity: $[\dot{x}, \dot{y}, \dot{z}]$
- Fingertip Position Covariance: $[triu(Log(C_p))]$
- Fingertip Velocity Covariance: $[triu(Log(C_v))]$

Tactile Information: A fingertip with a custom-made magnetic sensor, presented in [17], is attached to each finger of the AH. To quantify the intensity of force applied on each fingertip, we selected the norm of the tactile force as the feature. This allows us to accurately capture and analyze the dynamic interactions between the robot fingertips and manipulated objects. Additionally, a feature indicating the contact status is selected to reflect the interaction scenario between the AH and the object. This feature essentially distinguishes whether there is direct contact between the fingertip and the object. Incorporating this feature also aids in reducing noise, thereby enhancing the reliability and clarity of the sensory data used for task analysis and execution.

- Tactile Norm $(Fx, Fy, Fz) \times 4$
- Contact Status: $[(0, 1) \times 4]$

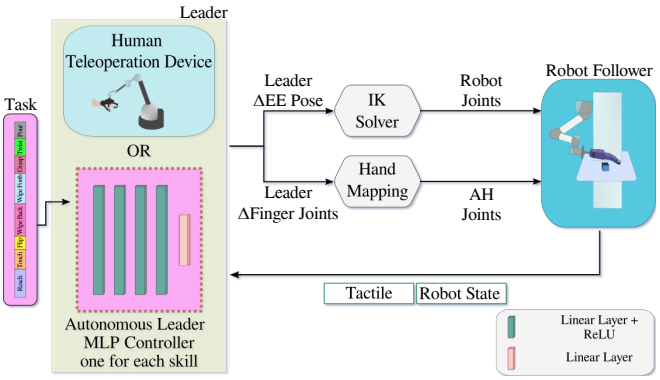


Fig. 2: The leader agent generates motor control commands for the end effector pose and finger joints of the hand. The follower robot executes corresponding actions based on these commands. During teleoperation, the follower robot provides haptic feedback. When operating the robot autonomously, we control the robot using a distinct MLP trained on the proprioceptive and tactile data (i.e. haptic data) of each separate skill.

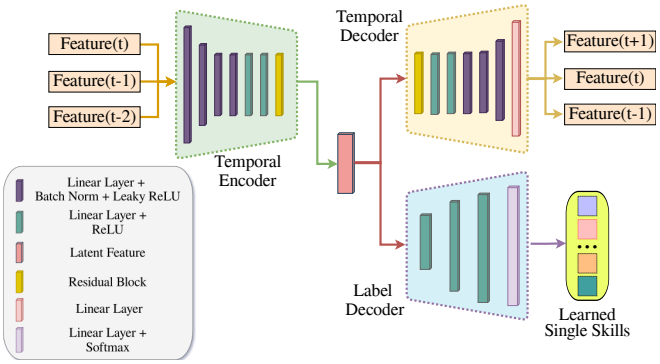


Fig. 3: The architecture of our Neural Network for supervised representation learning incorporates an auto-regressive autoencoder and a label decoder. This network processes time-series feature data as input, with the encoder transforming these features into a latent space. The temporal decoder reconstructs the features along with their predictions, whereas the label decoder extracts labels from the latent vectors. The label decoder is jointly trained with the autoencoder generating latent features that improve the segmentation performance.

B. Primitive Skills Learning

Long-horizon dexterous manipulation tasks often demand extensive demonstrations and face challenges related to accumulating errors and limited generalization capabilities. When learning long-horizon dexterous manipulation skills, humans typically analyze and arrange primitive skills in sequence. Motivated by this human approach, our study introduces

a methodology that first focuses on separately acquiring primitive skills and then predicting the combination sequence of these skills for new long-horizon tasks. The specific primitive skills we proposed for robotic manipulation using dexterous hands are detailed in Table I.

The process of collecting primitive skills datasets and training the autonomous skill policy is shown in Fig. 2. The learned manipulation policy operates with the assumption that the robot already knows the location of the object. For each primitive skill, an MLP neural network is utilized to learn the state-action pair from human demonstrations using the BC method. Given state information s, the goal of the BC is to train a policy $\pi(\theta_p)$ that can learn the behaviour of the human demonstrator by predicting the action a:

$$\theta_p = \arg\max_{\theta_p} \mathop{\mathbb{E}}_{\mathbf{s},\mathbf{a}\sim\mathcal{D}} [L(\mathbf{a} \mid \mathbf{s})] \tag{1}$$

where s is the input to the network, which includes the robot EE state, AH joint state, and filtered tactile information as well as the contact indicator. $L(\cdot)$ represents the loss function. In this work, the mean squared error (MSE) is used to calculate the loss between the predicted action and the recorded human action during the demonstration. The network output $a = [\Delta EE, \Delta Fingertip]$ captures the movement of both EE and AH fingertip, reflecting the actions executed by humans during the teleoperation of the robot system. This approach enables the policy $\pi(\theta_p)$ to directly learn from human input, controlling the robot through the same mapping to bypass the complexities of joint-space control and enhance safety.

C. Supervised Representation Learning

The proposed segmentation framework is shown in Fig. 3. This framework consists of two parts. Firstly, a temporal auto-regressive AE is utilized to encode temporal features into a latent space and then reconstruct the features for the subsequent time step. Secondly, a label decoder extracts the task label from the encoded latent representation. The AE receives an input sequence represented by $[\mathbf{f}_{t-2}, \mathbf{f}_{t-1}, \mathbf{f}_t]$, while the output generated by the encoder is $[\mathbf{f}_{t-1}, \mathbf{f}_t, \mathbf{f}_{t+1}]$, where f denotes the features outlined in Section III-A. The interval between each timestep in this process is set to 0.1 seconds. Furthermore, the representation of skill labels utilizes a one-hot encoding scheme.

The choice of an auto-regressive AE is strategic for encapsulating the temporal dynamics of robot behaviour, crucial for capturing the intricate nature of robot actions where the sequence of actions is significant. This setup is designed to capture the nuances of human behaviour through temporal data analysis. Additionally, the AE facilitates a deeper comprehension of the input data, which in turn, augments the accuracy of task label predictions by understanding the underlying patterns and characteristics of human behavior encoded in the temporal information. During the training, we train the AE together with the label decoder. The loss function is defined as:

$$\text{Loss} = \text{MSE}(f_{\text{true}}, f_{\text{pred}}) + \text{CrossEntropy}(l_{\text{true}}, l_{\text{pred}}) \tag{2}$$

where the loss $\text{MSE}(f_{\text{true}}, f_{\text{pred}})$ calculates the MSE between the true temporal features f_{true} and the predicted temporal features f_{pred}. The loss function $\text{CrossEntropy}(l_{\text{true}}, l_{\text{pred}})$ is specifically selected to calculate the discrepancy between the actual probabilities l_{true} and the predicted probabilities l_{pred} for each class. The combination of these two loss functions encourages the model to learn rich and informative information that helps in both accurately predicting temporal features and correctly identifying skill labels.

D. Long-horizon manipulation skills learning

Our objective is to facilitate one-shot imitation learning for long-horizon dexterous manipulation tasks that were previously unseen, utilizing a set of predefined primitive skills. Segmenting the long-horizon tasks into primitive skills makes it more structured for the algorithm to master the task, reducing the overall complexity. The trained feature encoder $\mathcal{E}$ and the label decoder $\mathcal{L}$ are used to predict the sequence of primitive skill for executing dexterous long-horizon manipulation tasks:

$$\mathbf{P_t} = \mathcal{L}(\mathcal{E}(\mathbf{f}_{t-2}, \mathbf{f}_{t-1}, \mathbf{f}_t; \theta_{\mathcal{E}}); \theta_{\mathcal{L}}), \tag{3}$$

where $\mathbf{P_t}$ is the predicted probability distribution of primitive skill at timestep t, $\theta_{\mathcal{E}}$ and $\theta_{\mathcal{L}}$ are the parameters of the feature encoder and label decoder respectively. To enhance the accuracy of these predictions and reduce errors, a median filter is applied for smoothing. After identifying a sequence of the primitive skills for these dexterous long-horizon tasks, the trained primitive skills are executed in sequence to complete the tasks.

IV. Experimental Setting

A. Teleoperation Setup

Depicted in Fig. 1, the teleoperation setup used is similar to the one proposed by the authors in a previous work [17]. In summary, the setup consists of two leader devices: Virtuose 6D and HGlove, used to respectively control the robotic arm UR5 and Allegro dexterous Hand mounted with custom-made magnetic sensors on the follower side. The mapping and control structure is the same as in the previous study, as referenced and summarized in Fig. 2. In this study, the wrist rotation of the robotic arm is enabled in order to achieve a wider range of skills.

For the data collection phase, all the information used to control the robot and to extract the training features proposed in Sec.III-A was recorded and made available in a dataset on the project's repository[1].

B. The Primitive Skills

In this work, amongst the various possible robotics manipulation skills, 20 primitive skills have been identified and are listed with a unique colour associated in Table I. A combination of more than three skills defines a Long-Horizon task. Although all skills are shown independently, some of them require a pre-action in the real world to be performed. The task of planning long-horizon tasks is left to the demonstrator. The objective of this work is to provide

TABLE I: List of primitive skills

Group	Skills	
	Pre	Skill
No Touch	-	1. Reach
	-	2. Setup Position
Touch	3. PreTouch	4. Touch
	-	5. Flip
	PreTouch	6. Wipe Forth
	PreTouch	7. Wipe Back
	8. PreGrasp	9. Grasp
	-	10. Lift with Grasp
	-	11. Transport Forward
	-	12. Place
	13. PreRotate	14. Rotate
	-	15. Shake Up
	-	16. Shake Down
	-	17. Twist
	-	18. Vertical Place
	-	19. Pour
	-	20. Release

TABLE II: Long-Horizon tasks: primitive skill recombinations. Tasks are denoted by alphabetical letters ranging from A to T. Objects utilized for the demonstrations are indicated within parentheses: (s) sponge, (t) tomato passata package, and (b) bottle containing liquid.

Task	Skills									
	I	II	III	IV	V	VI	VII	VIII	IX	X
A (s)	1	5	3	4	7	6	8	9	10	20
B (t)	4	7	8	9	10	11	12	2		
C (b)	13	14	10	15	16	17	18			
D (s)	6	7	6	7	6	7				
E (b)	5	8	9	10	15	19				
F (b)	8	9	10	17						
G (b)	1	5	8	9						
H (t)	15	16	15	12						
I (s)	16	15	16	20						
J (b)	9	10	17	20						
K (t)	4	8	9							
L (s)	13	14	17							
M (s)	9	20	2							
N (s)	17	10	16							
O (b)	10	17	19							
P (t)	19	17	18							
Q (s)	5	8	2							
R (b)	1	13	2							
S (s)	18	10	20							
T (b)	10	17	18							

a paradigm for segmentation and learning of skills where the feasibility of a task is defined by the demonstration sequences. For this reason, the assumption is made that each task is demonstrable and repeatable in a real-world scenario.

C. Long-Horizon Tasks

As the number of skills increases, the combinations grow exponentially. Although it is possible to have a sequence with many different skills that can turn into a challenging demonstration, even a basic task repeated multiple times may involve just three skills, . For instance, a sequence comprising ”Wipe Forth” and ”Wipe Back” combined with other skills like ”Pick” and ”Place” or ”Twist” may resemble a dish-washing operation. Similarly, an experiment involving ”Shake Down,” ”Shake Up,” and ”Pour” might evoke the actions of mixing and serving a cocktail.

For this study, we have identified 20 sparse long-horizon tasks of varying complexity, as outlined in Table II, which define the experimental set for the classifier. In this table, each long-horizon task is represented by a numbered skill sequence and with the same colours according to Table I. To illustrate the independence of skills from each other in skill recombination, it has been occasionally chosen to initiate from pre-established verified conditions, as observed in tasks B, D, J, K, and M. We will evaluate the performance of our framework on these combinations in Section VI-A.

In the experiments section, we will replicate autonomously long-horizon tasks defined by sequences A and B. As detailed in the section VI-B, this will be achieved from learning from the single skills demonstrations used for classifier training. Thus, no long-horizon sequences will be used in any of the training phases.

V. Data Collection

This section delineates the data collection methodology used to evaluate the proposed approach, while Section VI offers a discussion of the achieved results. For all the experiments, it is assumed that the position of the object is known.

The data collection is organized in three phases:

A. The training data for the 20 primitive skills were gathered through human teleoperation of the robot, with the majority of skills documented across 10 demonstrations involving a single object. Specifically, for the skills ”Touch,” ”Wipe Back,” and ”Grasp,” data were collected using both a soft sponge (16x8.5x4.5cm, 20g) and cardboard packages (14x7.5x4 cm, 15g), with 20 demonstrations performed for each skill. Additionally, for the skill ”Lift with Grasp,” an extra 10 demonstrations were conducted using a bottle (22x7x7cm, 220g). This collected data was then used to train the model for the autonomous primitive skill classifier, as shown in Figure 3.

B. The first phase of experiments is made to evaluate the performance of the classifier. The demonstrations of the 20 long-horizon tasks presented in Table II were collected. Each demonstration of a long-horizon task was presented and manually segmented with labels to make a comparison with the predictions obtained by the classifier. This process is depicted in Fig. 1.

C. The second set of experiments consists of replacing the Leader agent with distinct MLP controllers shown in Fig. 2, each tailored to a specific skill, trained from the same demonstrations used for the classifier. As in the methods section III-D, on receiving a skill label, the robot reproduces the single learned skill and switches to the next one on receipt of a new skill label.

All collected data are organized into datasets for future use. The dataset includes 200 trajectories of primitive skills gathered from human demonstrations via teleoperation and 20 long-horizon task trajectories, each consisting of various combinations of primitive skills.

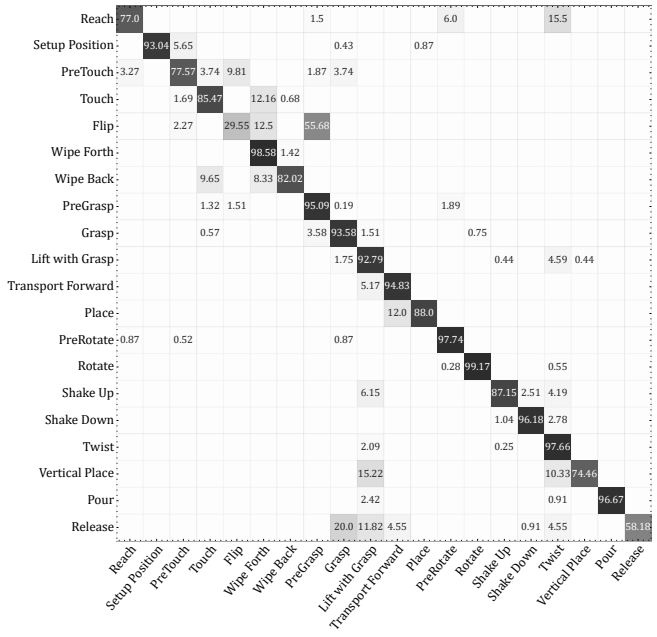


Fig. 4: Confusion matrix (%) of the segmentation system on the Long-horizon demonstrations (detailed in Section IV-C).

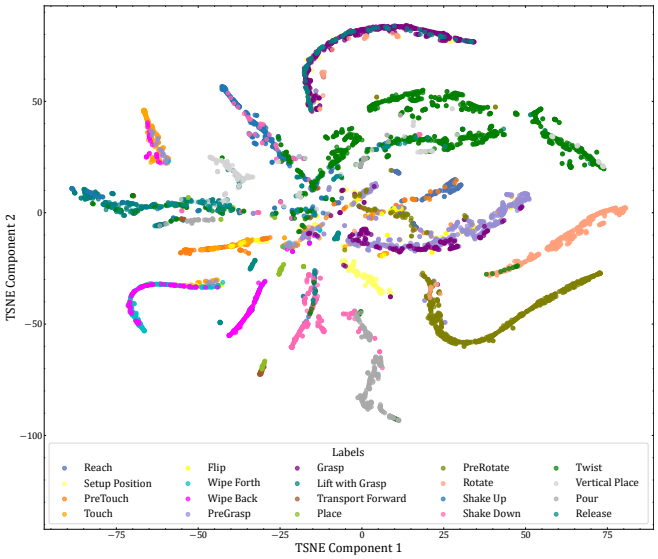


Fig. 5: T-SNE visualization of the classifier latent features. Each point in the graph corresponds to a primitive skill instance, differentiated by various colours to distinguish among the primitive skills.

VI. Experimental Results

In this section, we present the evaluation of our framework through a series of experiments. We explore the impact of integrating an auto-regressive AE on the accuracy of skill label prediction. Additionally, we investigate the benefits of jointly training the auto-regressive AE with the label decoder for overall performance improvement. Our experiments also validate the efficiency of the proposed haptic features in accurately differentiating between primitive skills. Finally, we demonstrate the ability of our framework to efficiently select and integrate skill primitives for the execution of complex, long-horizon manipulation tasks.

A. Classifier Evaluation

We train the classifier using data collected from primitive skills and validate its performance on unseen long-horizon manipulation tasks. The training is completed after 500 epochs, lasting about 40 minutes. We select the model with the lowest training loss from all the epochs. To validate the effectiveness of our framework, we chose parameters including Accuracy, Precision, Recall, F1-score, and Average IoU as our key performance indicators.

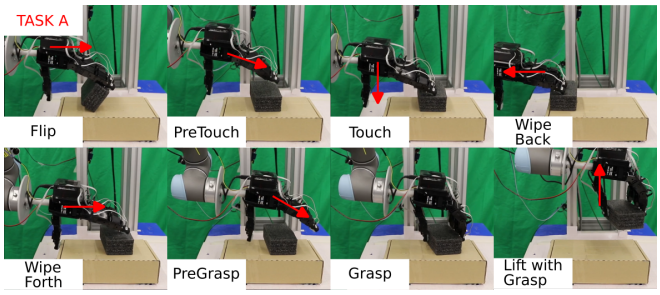
1) Framework Validation: We validate the performance of our proposed framework for long-horizon dexterous manipulation task segmentation by comparing it against three baselines. First, we examine a setup that utilizes a pre-trained AE for feature encoding, which is then followed by training a label decoder on these encoded features. Then, our second baseline explores the combination of a feature recovery AE with a label decoder, thereby excluding the temporal prediction from the AE. Finally, we verified the effectiveness of training the feature encoder and the label decoder together, excluding the temporal feature decoder. The outcomes are presented in Table III, where "LD" denote the label decoder. Our framework achieved an accuracy of 91%, outperforming all baseline comparisons. This indicates that jointly training the temporal AE with the label decoder is beneficial for capturing the latent dynamics of robot behaviour while preserving essential information critical for differentiating between primitive skills. This capability significantly contributes to improving the performance of the classifier by ensuring a deeper comprehension of the temporal features and movements associated with various primitive skills.

2) Feature importance verification: To verify the importance of the proposed feature for skill classification, we evaluated the performance of our proposed framework in two distinct scenarios: firstly, the framework was trained on raw haptic data, including the end-effector state, filtered tactile information, filtered contact indicators, and the AH joint state, which are the data used in training the policy for primitive skills; secondly, by training only with the proposed features while excluding the raw haptic data. The results are shown in Table III. When trained with raw haptic information, the classifier achieves an accuracy of only 35%. In contrast, utilizing the additional features proposed for skill segmentation improves the accuracy to approximately 76%. This substantial improvement indicates the effectiveness of our proposed features for skill segmentation, highlighting their value in enhancing the performance of the classifier.

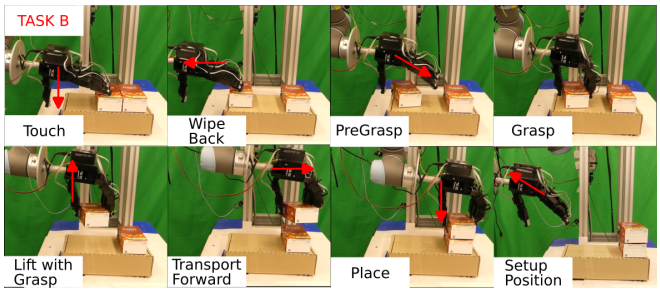
3) Confusion matrix for each skill: The confusion matrix provides a comprehensive overview of how well a classification model is performing across all categories. The result of the confusion matrix on 20 unseen long-horizon tasks is shown in Fig. 4. The prediction outcomes for the majority of the skills are satisfactory. However, the skills of flip and release tend to be confused with other skills, primarily due to their short duration, which makes their

TABLE III: Comparative Performance Evaluation

	DexSkills	Pre-trained AE+ LD	Feature recovery AE + LD	Feature Encoder + LD	Raw haptic data	Calculated feature
Accuracy	0.91	0.83	0.73	0.83	0.35	0.76
Precision	0.89	0.81	0.64	0.80	0.40	0.77
Recall	0.85	0.76	0.66	0.76	0.37	0.69
F1-Score	0.87	0.76	0.63	0.75	0.34	0.69
Average IoU	0.79	0.64	0.56	0.63	0.23	0.55



(a) Long-horizon Task A using a soft sponge (16x8.5x4.5cm, 20g).



(b) Long-horizon Task B using a cardboard package (14x7.5x4 cm, 15g).

Fig. 6: Autonomous skills reproduction. Full videos are available on the project repository

features challenging to capture accurately. Additionally, it shall be noted that manual labelling of the ground truth in long-horizon tasks can introduce certain errors inevitably, which can become especially obvious during the transition phases between skills.

B. Autonomous Long-horizon Tasks Execution

In this work, we verified our framework on two challenge tasks, which are identified as tasks A and B in Table II. Each task required the coordination of over 8 distinct skills, with skill transitions manually controlled by a human operator. Both tasks were completed by using the predicted primitive skills combination sequence. This demonstrates the ability of our proposed framework to reuse primitive skills for achieving unseen long-horizon tasks effectively. The snapshots of these experiments are presented in Figures 6a and 6b. Additionally, videos detailing the experimental process are available on the project GitHub page[1].

C. T-SNE Analysis

We employed t-Distributed Stochastic Neighbor Embedding (t-SNE) analysis on the latent features extracted before the final layer of the label decoder. This technique projects high-dimensional vectors onto a two-dimensional plane, achieving the visualization of high-dimensional data relationships. The application of t-SNE to our unseen long-horizon task data is depicted in Fig. 5, where a distinct colour represents each skill.

Overall, the t-SNE plot demonstrates that most skills are well-clustered and exhibit clear distinctions from one another. However, for skills that exhibit similar behaviours or tend to occur sequentially such as ”PreGrasp”, ”Grasp”, and ”Lift with Grasp”, there is a noticeable overlap in their representation. This overlapping tendency is also observed among skills with similar motions, like ”Wipe Back” and ”Wipe Forth”.

VII. Conclusion

In this study, we have presented DexSkills a learning framework that addresses the challenge of executing real-world long-horizon tasks with dexterous robotic hands by decomposing them into reusable primitive skills, trained from human demonstrations. This framework, adaptable for various demonstrations such as bilateral or kinesthetic teleoperation, enables the segmentation of long-horizon demonstrations into individual tasks, using only proprioceptive and tactile data, facilitating their integration into the repertoire of known skills and enabling autonomous robot execution of diverse tasks with high classification accuracy. In this study, we evaluate the segmentation performance of unseen multiple long-horizon tasks starting from a limited set of primitive skills. Our framework demonstrated strong performance in skill segmentation, achieving an accuracy rate of 91%. Comparative analysis reveals that our framework surpasses alternative approaches by effectively capturing the latent dynamics of primitive skills within the feature set. Furthermore, through the same demonstrations employed for the Classifier, a protocol is presented for learning the autonomous control over each individual task. Although the accuracy was not studied, it is established how each of the demonstrated tasks can be replicated autonomously by the robot by providing a label sequence. Future work should include the development of more complex models for autonomous reproduction and should accurately assess the reproducibility of autonomous execution of long-horizon tasks. Additionally, developing an effective representation learning framework that enables the incorporation of new primitive skills without retraining the entire framework from scratch is also an interesting topic to enhance adaptability.

REFERENCES

[1] K. Yao and A. Billard, "Exploiting kinematic redundancy for robotic grasping of multiple objects," *IEEE Transactions on Robotics*, 2023.

[2] X. Gao, K. Yao, F. Khadivar, and A. Billard, "Real-time motion planning for in-hand manipulation with a multi-fingered hand," *arXiv preprint arXiv:2309.06955*, 2023.

[3] F. Khadivar and A. Billard, "Adaptive fingers coordination for robust grasp and in-hand manipulation under disturbances and unknown dynamics," *IEEE Transactions on Robotics*, 2023.

[4] S. Wang, W. Hu, L. Sun, X. Wang, and Z. Li, "Learning adaptive grasping from human demonstrations," *IEEE/ASME Transactions on Mechatronics*, vol. 27, no. 5, pp. 3865–3873, 2022.

[5] A. Petrenko, A. Allshire, G. State, A. Handa, and V. Makoviychuk, "Dexpbt: Scaling up dexterous manipulation for hand-arm systems with population based training," *arXiv preprint arXiv:2305.12127*, 2023.

[6] W. Hu, F. Acero, E. Triantafyllidis, Z. Liu, and Z. Li, "Modular neural network policies for learning in-flight object catching with a robot hand-arm system," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 944–951.

[7] A. Handa, A. Allshire, V. Makoviychuk, A. Petrenko, R. Singh, J. Liu, D. Makoviichuk, K. Van Wyk, A. Zhurkevich, B. Sundaralingam *et al.*, "Dextreme: Transfer of agile in-hand manipulation from simulation to reality," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 5977–5984.

[8] S. Zhaole, J. Zhu, and R. B. Fisher, "Dexdlo: Learning goal-conditioned dexterous policy for dynamic manipulation of deformable linear objects," *arXiv preprint arXiv:2312.15204*, 2023.

[9] C. Yu and P. Wang, "Dexterous manipulation for multi-fingered robotic hands with reinforcement learning: a review," *Frontiers in Neurorobotics*, vol. 16, p. 861825, 2022.

[10] S. Luo, J. Bimbo, R. Dahiya, and H. Liu, "Robotic tactile perception of object properties: A review," *Mechatronics*, vol. 48, pp. 54–67, 2017.

[11] Q. Li, O. Kroemer, Z. Su, F. F. Veiga, M. Kaboli, and H. J. Ritter, "A review of tactile information: Perception and action through touch," *IEEE Transactions on Robotics*, vol. 36, no. 6, pp. 1619–1634, 2020.

[12] J. Jiang, G. Cao, J. Deng, T.-T. Do, and S. Luo, "Robotic perception of transparent objects: A review," *IEEE Transactions on Artificial Intelligence*, 2023.

[13] Z.-H. Yin, B. Huang, Y. Qin, Q. Chen, and X. Wang, "Rotating without seeing: Towards in-hand dexterity through touch," *arXiv preprint arXiv:2303.10880*, 2023.

[14] K.-W. Lee, Y. Qin, X. Wang, and S.-C. Lim, "Dextouch: Learning to seek and manipulate objects with tactile dexterity," *arXiv preprint arXiv:2401.12496*, 2024.

[15] S. Pai, T. Chen, M. Tippur, E. Adelson, A. Gupta, and P. Agrawal, "Tactofind: A tactile only system for object retrieval," *arXiv preprint arXiv:2303.13482*, 2023.

[16] W. Hu, B. Huang, W. W. Lee, S. Yang, Y. Zheng, and Z. Li, "Dexterous in-hand manipulation of slender cylindrical objects through deep reinforcement learning with tactile sensing," *arXiv preprint arXiv:2304.05141*, 2023.

[17] G. Giudici, B. Omarali, A. A. Bonzini, K. Althoefer, I. Farkhatdinov, and L. Jamone, "Feeling good: Validation of bilateral tactile telemanipulation for a dexterous robot," in *Annual Conference Towards Autonomous Robotic Systems*. Springer, 2023, pp. 443–454.

[18] Í. Elguea-Aguinaco, A. Serrano-Muñoz, D. Chrysostomou, I. Inziarte-Hidalgo, S. Bøgh, and N. Arana-Arexolaleiba, "A review on reinforcement learning for contact-rich robotic manipulation tasks," *Robotics and Computer-Integrated Manufacturing*, vol. 81, p. 102517, 2023.

[19] A. Correia and L. A. Alexandre, "A survey of demonstration learning," *arXiv preprint arXiv:2303.11191*, 2023.

[20] J. Luo, C. Xu, X. Geng, G. Feng, K. Fang, L. Tan, S. Schaal, and S. Levine, "Multi-stage cable routing through hierarchical imitation learning," *IEEE Transactions on Robotics*, 2024.

[21] L. X. Shi, A. Sharma, T. Z. Zhao, and C. Finn, "Waypoint-based imitation learning for robotic manipulation," *arXiv preprint arXiv:2307.14326*, 2023.

[22] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, p. eaat8414, 2019.

[23] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent advances in robot learning from demonstration," *Annual review of control, robotics, and autonomous systems*, vol. 3, pp. 297–330, 2020.

[24] V. Kumar, A. Gupta, E. Todorov, and S. Levine, "Learning dexterous manipulation policies from experience and imitation," *arXiv preprint arXiv:1611.05095*, 2016.

[25] I. Radosavovic, X. Wang, L. Pinto, and J. Malik, "State-only imitation learning for dexterous manipulation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 7865–7871.

[26] S. P. Arunachalam, S. Silwal, B. Evans, and L. Pinto, "Dexterous imitation made easy: A learning-based framework for efficient dexterous manipulation," in *2023 ieee international conference on robotics and automation (icra)*. IEEE, 2023, pp. 5954–5961.

[27] E. Valassakis, G. Papagiannis, N. Di Palo, and E. Johns, "Demonstrate once, imitate immediately (dome): Learning visual servoing for one-shot imitation learning," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 8614–8621.

[28] D. Zhang, Q. Li, Y. Zheng, L. Wei, D. Zhang, and Z. Zhang, "Explainable hierarchical imitation learning for robotic drink pouring," *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 4, pp. 3871–3887, 2021.

[29] F. Xie, A. Chowdhury, M. De Paolis Kaluza, L. Zhao, L. Wong, and R. Yu, "Deep imitation learning for bimanual robotic manipulation," *Advances in neural information processing systems*, vol. 33, pp. 2327–2337, 2020.

[30] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[31] I. Guzey, B. Evans, S. Chintala, and L. Pinto, "Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play," *arXiv preprint arXiv:2303.12076*, 2023.

[32] T. P. Tomo, M. Regoli, A. Schmitz, L. Natale, H. Kristanto, S. Somlor, L. Jamone, G. Metta, and S. Sugano, "A new silicone structure for uskin—a soft, distributed, digital 3-axis skin sensor and its integration on the humanoid robot icub," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2584–2591, 2018.

[33] C. Coppola and L. Jamone, "Master of puppets: Multi-modal robot activity segmentation from teleoperated demonstrations," in *2022 IEEE International Conference on Development and Learning (ICDL)*. IEEE, 2022, pp. 88–94.

[34] S. Pirk, K. Hausman, A. Toshev, and M. Khansari, "Modeling long-horizon tasks as sequential interaction landscapes," *arXiv preprint arXiv:2006.04843*, 2020.

[35] S. Krishnan, A. Garg, S. Patil, C. Lea, G. Hager, P. Abbeel, and K. Goldberg, "Transition state clustering: Unsupervised surgical trajectory segmentation for robot learning," *The International journal of robotics research*, vol. 36, no. 13-14, pp. 1595–1618, 2017.

[36] D. Meli and P. Fiorini, "Unsupervised identification of surgical robotic actions from small non-homogeneous datasets," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8205–8212, 2021.

[37] Y. Zhu, P. Stone, and Y. Zhu, "Bottom-up skill discovery from unsegmented demonstrations for long-horizon robot manipulation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4126–4133, 2022.

[38] W. Wan, Y. Zhu, R. Shah, and Y. Zhu, "Lotus: Continual imitation learning for robot manipulation through unsupervised skill discovery," *arXiv preprint arXiv:2311.02058*, 2023.

[39] S. Park, K. Lee, Y. Lee, and P. Abbeel, "Controllability-aware unsupervised skill discovery," *arXiv preprint arXiv:2302.05103*, 2023.

[40] H. Bharadhwaj, J. Vakil, M. Sharma, A. Gupta, S. Tulsiani, and V. Kumar, "Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking," *arXiv preprint arXiv:2309.01918*, 2023.

[41] E. Triantafyllidis, F. Acero, Z. Liu, and Z. Li, "Hybrid hierarchical learning for solving complex sequential tasks using the robotic manipulation network roman," *Nature Machine Intelligence*, vol. 5, no. 9, pp. 991–1005, 2023.

[42] Y. Chen, C. Wang, L. Fei-Fei, and C. K. Liu, "Sequential dexterity: Chaining dexterous policies for long-horizon manipulation," *arXiv preprint arXiv:2309.00987*, 2023.