APPEARED IN
Proc. IEEE Int. Conf. on Robotics & Automation, 2025

General-purpose Clothes Manipulation with Semantic Keypoints

Yuhong Deng[1], David Hsu[1,2]

***Abstract*— Clothes manipulation is a critical capability for household robots; yet, existing methods are often confined to specific tasks, such as folding or flattening, due to the complex high-dimensional geometry of deformable fabric. This paper presents *CLothes mAnipulation with Semantic keyPoints* (CLASP) for *general-purpose* clothes manipulation, which enables the robot to perform diverse manipulation tasks over different types of clothes. The key idea of CLASP is semantic keypoints—e.g., "right shoulder", "left sleeve", etc.—a sparse spatial-semantic representation that is salient for both perception and action. Semantic keypoints of clothes can be effectively extracted from depth images and are sufficient to represent a broad range of clothes manipulation policies. CLASP leverages semantic keypoints to bridge LLM-powered task planning and low-level action execution in a two-level hierarchy. Extensive simulation experiments show that CLASP outperforms baseline methods across diverse clothes types in both seen and unseen tasks. Further, experiments with a Kinova dual-arm system on four distinct tasks—folding, flattening, hanging, and placing—confirm CLASP's performance on a real robot.**

I. Introduction

Imagine an intelligent household robot taking care of our laundry chores. Toward this goal, we require a general-purpose robot to perform a broad range of clothes manipulation tasks, such as *"fold the T-shirt for storage"* and *"hang the skirt on the hanger"* (Fig. 1). Despite recent advances in learning clothes manipulation skills like folding and flattening [1], [2], these methods are limited to specific clothes and tasks. Clothes have high-dimensional state [3] and diverse geometric structures, which vary significantly across different categories. The complexity of clothes state and geometries makes it challenging to develop a general-purpose method for manipulating a wide variety of clothes in many different ways.

How can we find a state representation for general-purpose clothes manipulation? Considering that clothes have manually designed structures with significant geometric features like sleeves and shoulders, we adopt semantic keypoints on these features as the general spatial-semantic representation. Each semantic keypoint is represented by a language descriptor and its corresponding position. These keypoints, carrying semantic meaning, are relatively easy to extract from observations. Besides, semantic keypoints can (i) define where clothes can be manipulated, aiding high-level task planning, and (ii) capture the clothes' geometry, guiding low-level action execution.

To detect semantic keypoints under clothes' self-occlusion and deformation, we use a masked auto-encoder [4] to learn a spatiotemporal representation through reconstructing masked image sequences. Reconstruction requires the model to infer the geometric structure of clothes from partial observation. To utilize semantic keypoints in clothes manipulation, we develop a general-purpose CLothes mAnipulation method with Semantic keyPoints (CLASP). CLASP is a hierarchical learning method that integrates LLM-powered high-level task planning with low-level action execution based on semantic keypoints.

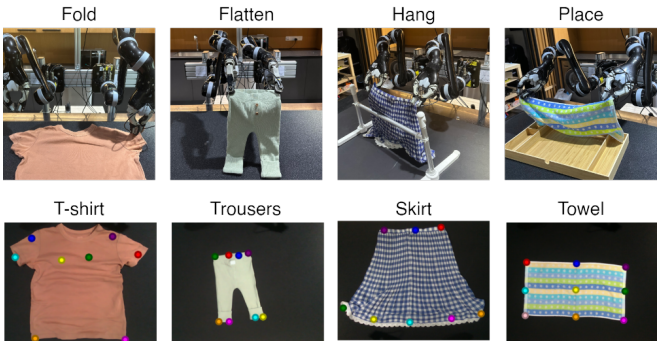

Fig. 1: General-purpose clothes manipulation. CLASP performs various manipulation tasks over different types of clothes.

For high-level task planning, we specify the goals for clothes manipulation using natural language instructions and then employ a large language model (LLM) for task planning. LLM provides commonsense knowledge for task planning, enhancing generalization [5]. Specifically, we prompt the LLM to decompose the given language instruction into sequential sub-tasks, where each sub-task is described by an action primitive and a contact point description. The contact point descriptions are selected from the language descriptors of semantic keypoints detected from the observation. Semantic keypoints can define where clothes can be manipulated in task planning. For low-level action execution, we establish a low-level action primitives library consisting of several heuristic policies parameterized by contact point positions. CLASP visually grounds contact points by retrieving from semantic keypoints based on their language descriptors. Based on the LLM proposed action primitive and contact point positions, we invoke a policy from a low-level action primitives library to generate the corresponding trajectory. Semantic keypoints serve as waypoints to guide trajectory generation.

To evaluate CLASP, we extend the SoftGym benchmark [6] to include 30 tasks across 4 common clothes categories and conduct simulation experiments. CLASP outperforms baseline methods in both seen and unseen tasks. Real-world experiments show that CLASP can be easily

[1] School of Computing, National University of Singapore, Singapore.{yuhongdeng, dyhsu}@comp.nus.edu.sg.
[2] Smart Systems Institute, National University of Singapore, Singapore.

transferred to the real world, performing well across a wide variety of clothes and tasks. These results demonstrate that semantic keypoints provide effective cues for clothes manipulation, enabling general-purpose clothes manipulation using our hierarchical learning method with LLM.

II. Related Work

A. Deformable Object Manipulation

There are two main methods for deformable object manipulation: model-based and data-driven methods. Particularly, model-based methods rely on a dynamic model to predict the configurations of deformable objects under a certain action and then select the appropriate action accordingly [7], [8]. However, deformable objects have complex non-linear dynamics, which are challenging to model. Recently, particle-based representation has become a unified solution for deformable object dynamics modeling. However, constructing particles from occluded deformable objects and accurate dynamics modeling is still challenging [9], [10].

Data-driven methods aim to learn robot actions directly from expert demonstrations without establishing a dynamic model. However, most data-driven methods are designed for specific tasks like rope rearrangement [11], cloth folding [12], cloth flattening [13], and bag opening [14]. Toward general-purpose deformable object manipulation, some goal-conditioned methods use goal images to specify different tasks for multi-task learning [15], [16]. However, goal-conditioned methods often struggle with generalization to unseen tasks. In this paper, we use semantic keypoints to represent the clothes state and propose a hierarchical learning method based on LLM, which performs effectively across a wide range of clothes categories and tasks, while also generalizing to unseen tasks.

B. Language-conditioned Object Manipulation

Language provides an intuitive interface in human-robot interaction and can explicitly capture the generalizable concepts between different manipulation tasks. Thus, language-conditioned object manipulation has been widely investigated. Early work focuses on making the robot understand language instructions and perform manipulation tasks [17], [18]. Recently, foundation models, such as large language models (LLMs) and vision-language models (VLMs), have been utilized in language-conditioned manipulation tasks. These models provide commonsense knowledge for reasoning [19], [20] and perception [21], [22], significantly enhancing generalization capabilities. Most of the previous language-conditioned object manipulation methods focus on rigid objects. However, a significant gap remains between task planning and action execution when it comes to deformable object manipulation. In this paper, we present semantic keypoints as an interface between LLM-powered task planning and low-level action execution, enabling general-purpose clothes manipulation.

C. State Representation of Deformable Object

Given the high-dimensional state of deformable objects, an effective state representation method is necessary. To simulate deformable objects, particles and mesh representations have been explored [23]–[25]. Such representation provides a solution to model deformable dynamics. However, they tend to be overly dense, making accurate state estimation challenging. Compared with particles and meshes, keypoints representation has lower dimensionality, leading to more effective policy learning [26]. Previous keypoints representation focuses on geometric shape, using keypoints to represent the topology of deformable object [27], [28]. In contrast, our semantic keypoints are represented by language descriptors and corresponding positions, effectively capturing both semantic and geometric information for manipulation tasks. As a result, our semantic keypoints can bridge and enhance both high-level task planning and low-level action execution.

III. Method

In this paper, we propose a general-purpose CLothes mAnipulation method with Semantic keyPoints (CLASP). The key idea of CLASP is using semantic keypoints as a general spatial-semantic representation of clothes. Each semantic keypoint is represented by a language descriptor like *"left sleeve"* and the corresponding keypoint position. The language descriptor can provide semantic information for task planning, while the keypoint position can provide geometry information to guide low-level action execution.

Fig. 2 illustrates the overall framework of CLASP. Given a depth image, a semantic keypoint detector will predict keypoint positions and corresponding language descriptors. The language descriptors of keypoints and the language instruction are fed into an LLM for task planning – inferring a sequence of sub-tasks, where each sub-task is described as an action primitive and a contact point description like `grasp`(*"left sleeve"*). For each sub-task, CLASP visually grounds contact points to pixel positions by retrieving corresponding semantic keypoints according to the contact point description. Finally, the contact point positions and action primitive will guide low-level action execution based on an action primitives library.

In the following sections, we will present how we detect effective semantic keypoints in Sec. III-A. Following that, we will introduce how we utilize semantic keypoints in high-level task planning (Sec. III-B) and low-level action execution (Sec. III-C) for general-purpose clothes manipulation.

A. Semantic Keypoint Detection

Although previous work has explored leveraging keypoints for object manipulation [22], robustly detecting semantic keypoints on clothes remains challenging, especially under self-occlusion and deformation. To address this, we utilize a mask autoencoder (MAE) [4] as a spatiotemporal learner to establish a powerful latent space. The core idea is that masking acts as a form of occlusion, and recovering the masked areas requires the model to infer geometric structures

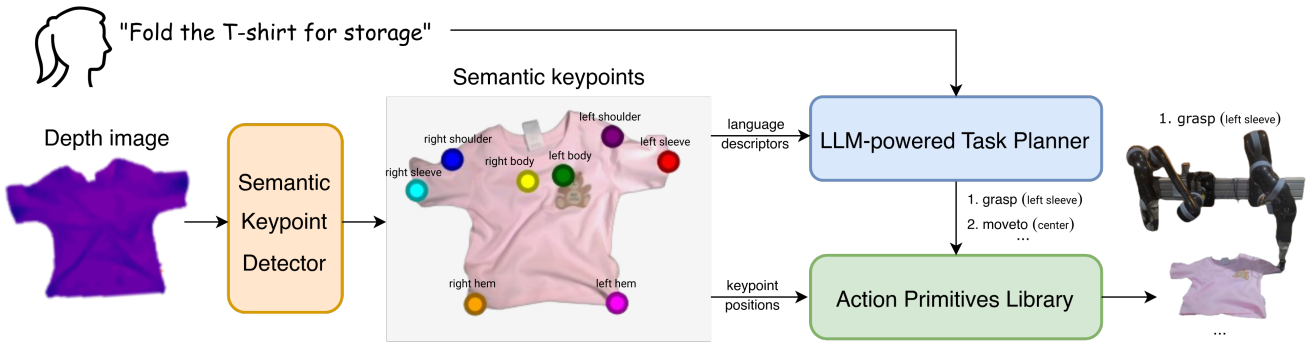

Fig. 2: An overview of CLASP. Given the natural-language task instruction and a depth image, CLASP first detects semantic keypoints, each consisting of a language descriptor and a 2-D geometric position. The task instruction and the language descriptors are fed into an LLM to generate a sequence of sub-tasks on the keypoints. For each sub-task, a low-level action primitives library generates the action on the keypoint position.

of clothes from partial observations. Specifically, we first establish a dataset $D = \{\zeta_1, \zeta_2, ..., \zeta_n\}$ of n sequences with associated discrete-time depth image, keypoints positions, and language descriptors pairs $\zeta_i = \{(I_t, P_t, S_t)\}_{t=1}^{T}$, where I_t refers to the depth image, P_t and S_t refer to the corresponding keypoints positions and language descriptors. We collect such a dataset in simulation through data argumentation on clothes' shape, pose, and configuration.

The training process of semantic keypoint detection consists of two stages: the reconstruction stage and the keypoint detection stage. In the first stage, we apply a random masking strategy on the depth image sequence, resulting in masked depth images $\{\tilde{I}_t\}_{t=1}^{T}$. The masked depth images will pass through an encoder-decoder structure to reconstruct the original depth image sequence. Given the reconstructed image sequence $\{\hat{I}_t\}_{t=1}^{T} = f_{dec}(f_{enc}(\{\tilde{I}_t\}_{t=1}^{T}))$, we optimize the encoder f_{enc} and decoder f_{dec} by minimizing the reconstruction loss, $L_r = \sum_{t=1}^{T} \|\hat{I}_t - I_t\|$. In the second stage, we fine-tune the encoder f_{enc} with an additional keypoint decoder f_{kp} to predict the semantic keypoints heatmap $\hat{\mathcal{H}}_t = f_{kp}(f_{enc}(I_t))$. The position of the k-th keypoint, p_t^k, can be determined by $p_t^k = \text{argmax}\,\hat{\mathcal{H}}_t^k$, where $\hat{\mathcal{H}}_t^k$ corresponds to the k-th channel of the semantic keypoints heatmap. For each clothes category, we predefine a set of language descriptors, denoted as $\mathcal{S}$. Given the clothes category, the language descriptor of the k-th keypoint, s_t^k, is obtained by $s_t^k = \mathcal{S}_k$. Each channel of the semantic keypoints heatmap $\hat{\mathcal{H}}_t$ is responsible for detecting keypoints with the same language descriptor. Using the ground truth keypoint positions and their corresponding language descriptors, we generate the ground truth semantic keypoints heatmap $\mathcal{H}_t$. During the second stage, the encoder f_{enc} and keypoint decoder f_{kp} are optimized by minimizing the semantic keypoints heatmap prediction loss, $L_{kp} = \|\hat{\mathcal{H}}_t - \mathcal{H}_t\|$.

B. Task Planning

For general-purpose clothes manipulation, we utilize an LLM for task planning due to its powerful commonsense knowledge from extensive internet-scale data. We first use the LLM with a chain of thought prompting [29] to define action primitives for clothes manipulation. The LLM is prompted to (i) provide examples of clothes manipulation tasks, (ii) decompose these examples into basic actions, and (iii) summarize the actions used in step (ii) to identify action primitives. In this way, we identify the action primitives set, including `grasp`, `moveto`, `press`, `release`, `rotate`, and `pull`. These action primitives reflect LLM's commonsense knowledge, enhancing task planning. To generate sub-tasks, we prompt the LLM with examples of language instructions paired with desirable sub-tasks sequences. Given a language instruction and language descriptors of semantic keypoints, the LLM generates sub-tasks by selecting action primitives from predefined action primitives set and contact points from language descriptors of semantic keypoints.

C. Action Execution

To complete each sub-task, we establish a low-level action primitives library. In the low-level action primitives library, there are some heuristic rules to create waypoints based on the contact point positions and selected action primitives. Once the waypoints are determined, a motion planning algorithm is applied to generate the complete trajectory for the robot's execution. Specifically, the action primitive `grasp` picks up clothes parts through planar grasping with one or both arms. The action primitive `moveto` transports the clothes to a target position, and target positions are determined by semantic keypoints or the location of the receptacle (e.g., a hanger or a box), depending on the task context. The action primitive `press` smooths out wrinkles by pressing along the normal vector of the table. The pressing distance is fixed at 1 cm in this paper. The action primitive `release` drops clothes parts by opening grippers. The action primitive `rotate` rotates the clothes grasped by grippers. This primitive is frequently used to align the clothes with the hanger in hanging tasks, and the hanger's pose can be obtained using an open-vocabulary object detector. The action primitive `pull` stretches the clothes to flatten them. In this paper, we set the stretch distance to 10% of the distance between the grippers of two arms.

TABLE I: Semantic keypoint detection performance.

	AKD (pixel) ↓	AP_8 (%) ↑	AP_4 (%) ↑	AP_2 (%) ↑
DINOv2-S	16.3	63.3	39.7	15.6
DINOv2-B	12.6	68.8	42.5	16.9
DINOv2-L	12.1	70.2	42.5	17.0
DINOv2-g	10.0	75.2	47.5	18.7
MAE (from scratch)	5.3	84.4	66.2	41.1
Ours	**3.8**	**91.0**	**75.4**	**50.2**

IV. Experiments

We aim to answer the following research questions: (i) How well does our semantic keypoint detector detect effective semantic keypoints? (Sec. IV-A) (ii) Can CLASP perform well and generalize to a wide variety of clothes and manipulation tasks? (Sec. IV-B) (iii) Can CLASP trained in simulation be transferred to real-world clothes manipulation tasks? (Sec. IV-C)

A. Semantic Keypoint Detection Experiments

To evaluate the performance of our semantic keypoint detector, we compare it with several baseline methods. Baseline methods include DINOv2 [30] and MAE (from scratch). DINOv2 is a large pretrained vision model with powerful visual features. We directly use the features from DINOv2 and train a keypoint decoder to predict semantic keypoints. We evaluate four types of DINOv2 with increasing parameters: DINOv2-S (Small), DINOv2-B (Base), DINOv2-L (Large), and DINOv2-g (Giant). Larger models offer better performance but come with increased computational burden. MAE (from scratch) uses the same architecture as our method but is trained from scratch without a reconstruction stage. All baseline methods are trained for 200 epochs. Our model is trained for 100 epochs in both the reconstruction and keypoint detection stages. We use two standard metrics for keypoint detection: Average Keypoint Distance (AKD) and Average Precision (AP). AKD measures the average distance between ground truth and detected keypoints, while AP represents the proportion of correctly detected keypoints under the given threshold. Given that the observed depth images have a resolution of 224 × 224, we set thresholds at 8, 4, and 2 pixels.

The results are shown in TABLE I. Our method outperforms both baseline methods, demonstrating that the reconstruction stage effectively learns a powerful representation, which enhances the performance of semantic keypoint detection. To further analyze our method, we conduct an ablation study by varying the masking ratio and comparing training with image sequences versus a single image. The evaluation metrics include Average Keypoint Distance (AKD) and Mean Average Precision (MAP), where MAP represents the mean of average precision values calculated at different thresholds. As Fig. 3 shows, increasing the masking ratio allows the model to better infer the geometric structure of clothes under self-occlusion, leading to performance improvement on keypoint detection. The model achieves the best performance when the masking ratio reaches 0.75, beyond which the information becomes insufficient. Additionally, reconstructing image sequences yields better results than reconstructing a single image, as the temporal information helps the model capture consistent geometric features relevant to semantic keypoints. In summary, our model design enables the establishment of a powerful spatiotemporal representation, significantly improving the model's ability to detect semantic keypoints of clothes.

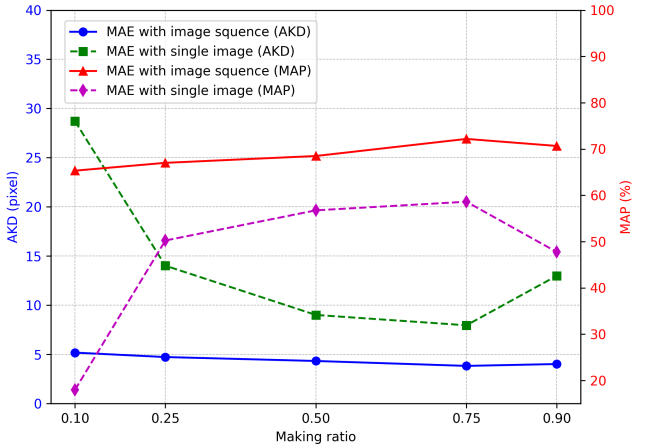

Fig. 3: An ablation study on the effects of masking and temporal information on semantic keypoint detection.

B. Simulation Experiments

To evaluate the proposed method's performance on clothes manipulation tasks, we conduct experiments in SoftGym [6], where clothes are modeled as particles with ground truth positions. 3D models of clothes are sampled from CLOTH3D [33] dataset, covering 4 common clothes categories in human's daily life: T-shirts, trousers, skirts, and towels. For each category, over 35 instances of varying sizes and shapes are used. In addition, we extend the SoftGym benchmark to 30 tasks. These tasks can be divided into 4 categories:

- **Folding** tasks involve folding clothes to achieve a target configuration. The success of a folding task is determined by the particle position error between the folded item and the target configuration.
- **Flattening** tasks involve flattening crumpled clothes with random deformations. The success of a flattening task is determined by the coverage area of the clothes.
- **Hanging** tasks require hanging clothes on a hanger. A hanging task is successful when the clothes are fully hung on the hanger without any part touching the ground.
- **Placing** tasks require placing the clothes in a box. A placing task is successful when the clothes are laid flat inside the box.

In our experimental setup, only half of tasks are seen during training through examples in the prompt or demonstrations. Unseen tasks involve new object categories and new requirements like folding direction, position, and times. For each task category and object category, we conduct 120 trials with different clothes configurations to calculate the

TABLE II: Simulation experiments. The average success rates (%) of CLASP and baseline methods on clothes manipulation.

Method	Flattening (unseen object)		Hanging (unseen object)		Placing (unseen object)		Folding (unseen object)		Folding (unseen requirement)		
	Trousers	Towel	T-shirt	Skirt	Trousers	T-shirt	Trousers	Skirt	Position	Direction	Times
CLIPORT [18]	8.3	9.2	76.7	66.7	70.0	73.3	0.0	0.0	76.7	65.0	0.0
Goal-conditioned Transporter [16]	10.0	6.7	36.7	40.0	36.7	60.0	0.0	0.0	8.3	0.0	0.0
FlingBot [31]	29.2	34.2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
FabricFlowNet [32]	N/A	N/A	N/A	N/A	N/A	N/A	0.0	2.5	30.0	3.3	0.0
CLASP	**50.0**	**55.0**	**93.3**	**73.3**	**93.3**	**93.3**	**86.7**	**83.3**	**100.0**	**96.7**	**95.0**

Method	Flattening (seen object)		Hanging (seen object)		Placing (seen object)		Folding (seen object)	
	T-shirt	Skirt	Trousers	Towel	Towel	Skirt	Towel	T-shirt
CLIPORT [18]	32.5	36.7	76.7	83.3	93.3	**93.3**	77.5	80.0
Goal-conditioned Transporter [16]	26.7	33.3	**100.0**	80.0	66.7	83.3	83.3	76.7
FlingBot [31]	**66.7**	**85.0**	N/A	N/A	N/A	N/A	N/A	N/A
FabricFlowNet [32]	N/A	N/A	N/A	N/A	N/A	N/A	93.7	**100.0**
CLASP	55.0	78.3	96.7	**96.7**	**96.7**	90.0	**100.0**	**100.0**

success rate. Baseline methods include two general multi-task learning frameworks (CLIPORT and Goal-conditioned Transporter) and two task-specific algorithms (FlingBot and FabricFlowNet):

- **CLIPORT** [18] represents a typical end-to-end algorithm for learning language-conditioned manipulation policies, which leverages a pretrained vision-language model.
- **Goal-conditioned Transporter** [16] represents a typical goal-conditioned transporter network for deformable object manipulation, which infers the optimal action from the current and goal images.
- **FlingBot** [31] is a self-supervised learning framework designed for flattening crumpled clothes using a fling action.
- **FabricFlowNet** [32] is a learning framework designed for clothes folding, which infers actions based on optical flow.

The experiment results are shown in TABLE II. Overall, our method outperforms the two multi-task learning methods on both seen and unseen tasks. Compared with Goal-conditioned Transporter, CLIPORT shows better generalization. Language instructions facilitate capturing similarities between different tasks and the pretrained vision-language model enables CLIPORT to capture such similarities. However, CLIPORT's generalization is limited when adapting to tasks with significantly different action sequences, such as transferring the skill from folding a T-shirt to folding trousers, because it learns task-specific manipulation policies in an end-to-end manner. In contrast, CLASP is a hierarchical learning framework that learns generalizable language and visual concepts across a wide range of clothes manipulation tasks. The commonsense knowledge from LLM allows CLASP to handle unseen tasks by decomposing them into predefined action primitives. Furthermore, semantic keypoints are task-agnostic, providing cues for task planning and action execution in unseen manipulation tasks.

Compared to the two task-specific algorithms, CLASP shows comparable performance on seen tasks, demonstrating the effectiveness of the proposed method for clothes manipulation. On unseen tasks, task-specific algorithms tend to fail due to distribution shifts. In contrast, CLASP performs well on unseen tasks.

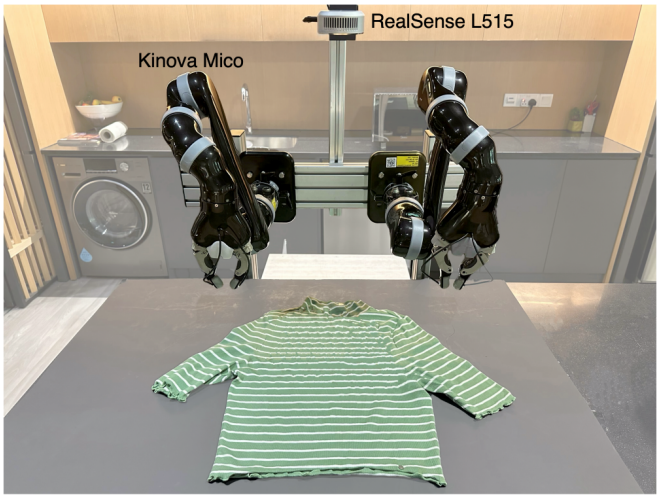

Fig. 4: Setup for real-robot experiments. The dual-arm system consists of an Intel RealSense camera for depth sensing and two Kinova Mico arms.

C. Real-Robot Experiments

We utilize depth images instead of RGB images to ensure our method can be directly transferred to real-world scenarios. To evaluate the sim-to-real performance of our method, we establish a dual-arm robot manipulation system. As Fig. 4 shows, the system consists of two Kinova Mico robot arms and a top-down RealSense L515 RGB-D camera for capturing depth images. The clothes are placed on a platform in front of the robot. To generate the dual-arm trajectories, we utilize a motion planning algorithm using MoveIt! [34] to avoid collisions and synchronize the motion between two arms. To evaluate the performance of our method, we select a diverse set of clothes varying in size, appearance, and shape, ranging from infant trousers to adult shorts.

We first evaluate the performance of our semantic keypoint detector, which is trained in simulation. To mitigate noise from depth sensing in real-world environments, we use

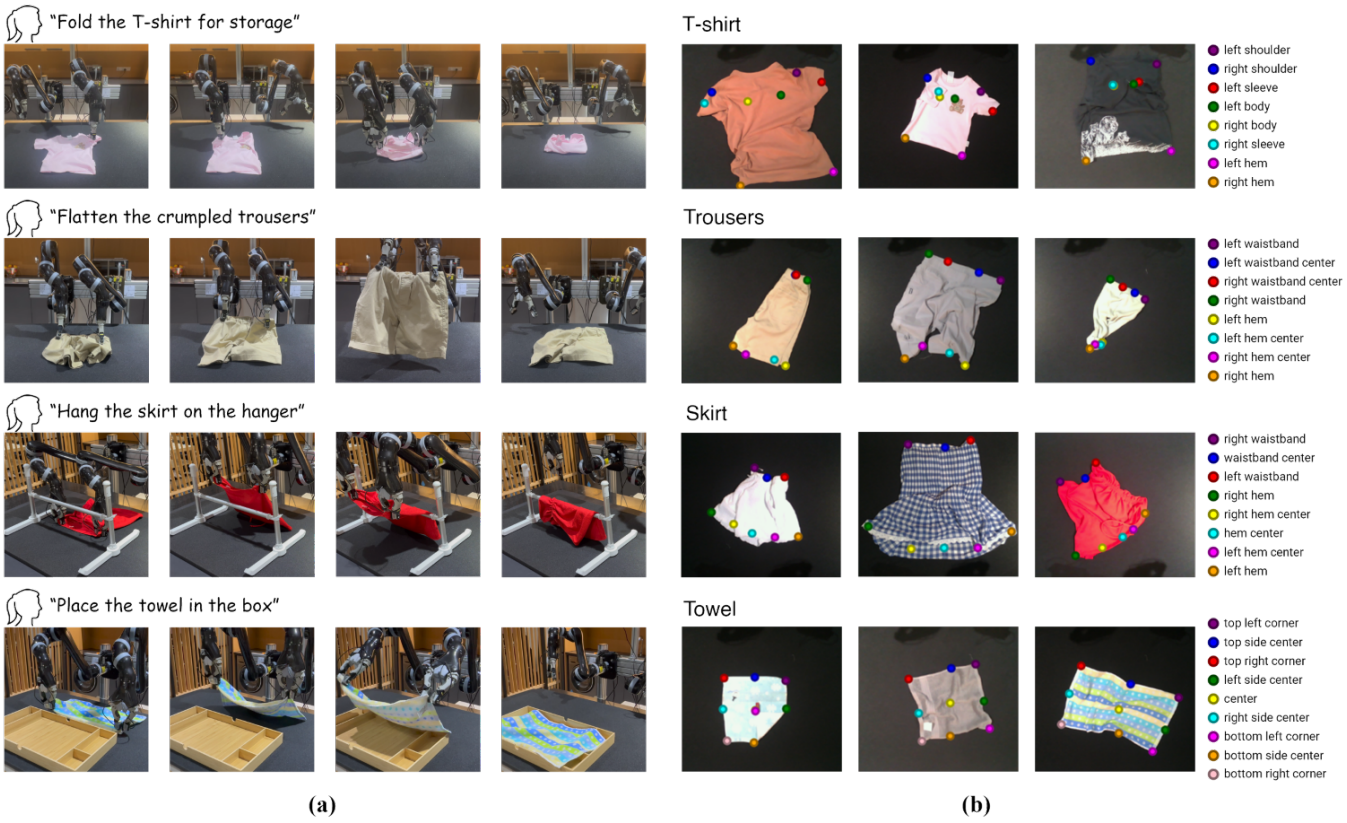

Fig. 5: Real-robot experiments. (a) Four clothes manipulation tasks: folding, flattening, hanging, and placing. (b) Semantic keypoint detection on a variety of clothes.

SAM [35] and OWLv2 [36] to generate masks that filter out noisy depth data and smooth the images. The experimental results are shown in Fig. 5 (b). Our semantic keypoint detector can robustly detect keypoints across a diverse range of clothes, even under irregular deformation and occlusion, without requiring fine-tuning on real-world images.

We further test the performance of CLASP on some real-world clothes manipulation tasks. Fig. 5 (a) illustrates four representative examples:*"Fold the T-shirt for storage"*, *"Flatten the crumpled trousers"*, *"Hang the skirt on the hanger"* and *"Place the towel in the box"*. CLASP demonstrates the ability to manipulate various clothes in many different ways in real-world scenarios.

V. Conclusion

In this paper, we propose semantic keypoints as a general spatial-semantic representation of clothes and enable general-purpose clothes manipulation using our hierarchical learning method with LLM. To detect semantic keypoints, we use a masked autoencoder to establish a spatiotemporal representation capable of handling self-occlusion and deformation. The integration of commonsense knowledge from the LLM and the general semantic keypoint representation ensures the generalization of our method. Simulation experiment results show that our method performs well on both seen and unseen tasks, including new object categories and task requirements. Furthermore, the proposed method can be directly transferred to real-world scenarios and performs well on a wide variety of clothes. However, our current solution is an open-loop system, where the task planning is performed only once at the beginning of the manipulation task, making it sensitive to unexpected state changes or external disturbances. Future improvement will focus on developing a closed-loop pipeline based on our semantic keypoints representation.

Acknowledgment

This research is supported by the National Research Foundation, Singapore, under its Medium Sized Centre Program, Center for Advanced Robotics Technology Innovation (CARTIN).

REFERENCES

[1] M. Moletta, M. K. Wozniak, M. C. Welle, and D. Kragic, "A virtual reality framework for human-robot collaboration in cloth folding," in *IEEE-RAS International Conference on Humanoid Robots*, 2023.

[2] A. Canberk, C. Chi, H. Ha, B. Burchfiel, E. Cousineau, S. Feng, and S. Song, "Cloth funnels: Canonicalized-alignment for multi-purpose garment manipulation," in *IEEE International Conference on Robotics and Automation*, 2023.

[3] H. Yin, A. Varava, and D. Kragic, "Modeling, learning, perception, and control methods for deformable object manipulation," *Science Robotics*, vol. 6, no. 54, p. eabd8803, 2021.

[4] C. Feichtenhofer, Y. Li, K. He *et al.*, "Masked autoencoders as spatiotemporal learners," *Advances in Neural Information Processing Systems*, 2022.

[5] Z. Zhao, W. S. Lee, and D. Hsu, "Large language models as commonsense knowledge for large-scale task planning," *Advances in Neural Information Processing Systems*, 2024.

[6] X. Lin, Y. Wang, J. Olkin, and D. Held, "Softgym: Benchmarking deep reinforcement learning for deformable object manipulation," in *Conference on Robot Learning*, 2021.

[7] Z. Hu, P. Sun, and J. Pan, "Three-dimensional deformable object manipulation using fast online gaussian process regression," *IEEE Robotics and Automation Letters*, 2018.

[8] M. Cusumano-Towner, A. Singh, S. Miller, J. F. O'Brien, and P. Abbeel, "Bringing clothing into desired configurations with limited perception," in *IEEE International Conference on Robotics and Automation*, 2011.

[9] K. Zhang, B. Li, K. Hauser, and Y. Li, "Adaptigraph: Material-adaptive graph-based neural dynamics for robotic manipulation," in *Proceedings of Robotics: Science and Systems*, 2024.

[10] S. Chen, Y. Xu, C. Yu, L. Li, and D. Hsu, "Differentiable particles for general-purpose deformable object manipulation," *arXiv preprint arXiv:2405.01044*, 2024.

[11] M. Yu, K. Lv, H. Zhong, S. Song, and X. Li, "Global model learning for large deformation control of elastic deformable linear objects: An efficient and adaptive approach," *IEEE Transactions on Robotics*, vol. 39, no. 1, pp. 417–436, 2022.

[12] K. Mo, C. Xia, X. Wang, Y. Deng, X. Gao, and B. Liang, "Foldsformer: Learning sequential multi-step cloth manipulation with space-time attention," *IEEE Robotics and Automation Letters*, 2022.

[13] Z. Xu, C. Chi, B. Burchfiel, E. Cousineau, S. Feng, and S. Song, "Dextairity: Deformable manipulation can be a breeze," in *Proceedings of Robotics: Science and Systems*, 2022.

[14] A. Bahety, S. Jain, H. Ha, N. Hager, B. Burchfiel, E. Cousineau, S. Feng, and S. Song, "Bag all you need: Learning a generalizable bagging strategy for heterogeneous objects," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2023.

[15] C. Chi, B. Burchfiel, E. Cousineau, S. Feng, and S. Song, "Iterative residual policy for goal-conditioned dynamic manipulation of deformable objects," in *Proceedings of Robotics: Science and Systems*, 2022.

[16] D. Seita, P. Florence, J. Tompson, E. Coumans, V. Sindhwani, K. Goldberg, and A. Zeng, "Learning to rearrange deformable cables, fabrics, and bags with goal-conditioned transporter networks," in *IEEE International Conference on Robotics and Automation*, 2021.

[17] M. Shridhar and D. Hsu, "Interactive visual grounding of referring expressions for human-robot interaction," in *Proceedings of Robotics: Science and Systems*, 2018.

[18] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Conference on robot learning*, 2022.

[19] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "Progprompt: Generating situated robot task plans using large language models," in *IEEE International Conference on Robotics and Automation*, 2023.

[20] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," in *IEEE International Conference on Robotics and Automation*, 2023.

[21] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "Voxposer: Composable 3d value maps for robotic manipulation with language models," in *Conference on Robot Learning*, 2023.

[22] W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei, "Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation," *arXiv preprint arXiv:2409.01652*, 2024.

[23] S. Chen, Y. Xu, C. Yu, L. Li, X. Ma, Z. Xu, and D. Hsu, "Daxbench: Benchmarking deformable object manipulation with differentiable physics," in *International Conference on Learning Representations*, 2022.

[24] J. Bender, M. Müller, and M. Macklin, "Position-based simulation methods in computer graphics." in *Eurographics (tutorials)*, 2015, p. 8.

[25] Y. Li, J. Wu, R. Tedrake, J. B. Tenenbaum, and A. Torralba, "Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids," in *International Conference on Learning Representations*, 2018.

[26] T. D. Kulkarni, A. Gupta, C. Ionescu, S. Borgeaud, M. Reynolds, A. Zisserman, and V. Mnih, "Unsupervised learning of object keypoints for perception and control," *Advances in Neural Information Processing Systems*, 2019.

[27] O. Gustavsson, T. Ziegler, M. C. Welle, J. Bütepage, A. Varava, and D. Kragic, "Cloth manipulation based on category classification and landmark detection," *International Journal of Advanced Robotic Systems*, vol. 19, no. 4, p. 17298806221110445, 2022.

[28] R. Shi, Z. Xue, Y. You, and C. Lu, "Skeleton merger: an unsupervised aligned keypoint detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[29] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, 2022.

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

[31] H. Ha and S. Song, "Flingbot: The unreasonable effectiveness of dynamic manipulation for cloth unfolding," in *Conference on Robot Learning*, 2022.

[32] T. Weng, S. M. Bajracharya, Y. Wang, K. Agrawal, and D. Held, "Fabricflownet: Bimanual cloth manipulation with a flow-based policy," in *Conference on Robot Learning*, 2022.

[33] H. Bertiche, M. Madadi, and S. Escalera, "Cloth3d: clothed 3d humans," in *European Conference on Computer Vision*, 2020.

[34] S. Chitta, I. Sucan, and S. Cousins, "Moveit![ros topics]," *IEEE Robotics & Automation Magazine*, vol. 19, no. 1, pp. 18–19, 2012.

[35] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[36] M. Minderer, A. Gritsenko, and N. Houlsby, "Scaling open-vocabulary object detection," *Advances in Neural Information Processing Systems*, 2024.