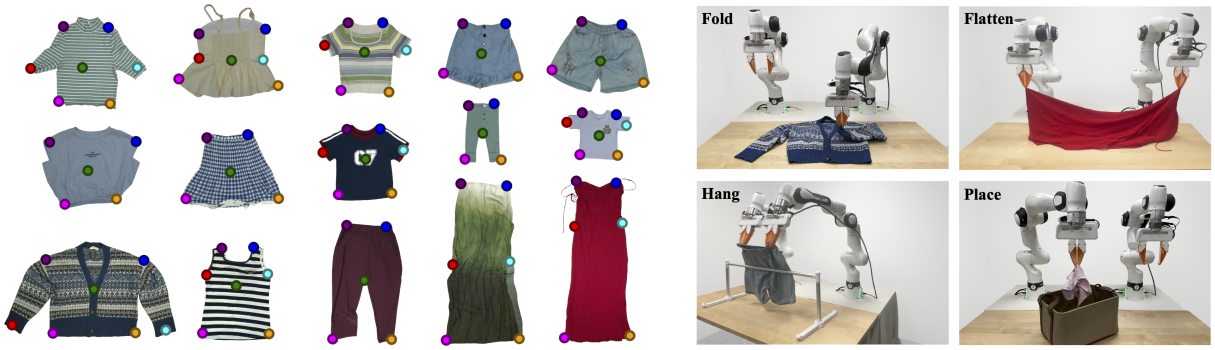
General-purpose Clothes Manipulation with Semantic Keypoints

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



(a) (b)

Fig. 1: CLothes mAnipulation with Semantic keyPoints (CLASP). The semantic keypoint representation enables CLASP to generalize over many different types of clothes and tasks. (a) Semantic keypoints for various types of clothes. (b) Four distinct clothes manipulation tasks.

***Abstract*— Clothes manipulation, such as folding and flattening, is a critical skill for home service robots. Despite recent advances, existing methods often focus narrowly on a specific task or a specific type of clothes. This work introduces *CLothes mAnipulation with Semantic keyPoints* (CLASP), which aims at general-purpose clothes manipulation over diverse clothes types and tasks. Our insight in tackling the challenge of generalization is the *semantic keypoints*, a general spatial-semantic representation that encodes the structural features of clothes, such as "left sleeve" and "right hem". Semantic keypoints are salient for both perception and action, effectively captured in the commonsense knowledge of foundation models, and relatively easy to extract from observations. CLASP integrates semantic keypoints with foundation models to achieve general-purpose clothes manipulation. In both simulation and real experiments, CLASP demonstrates strong performance and generalization capabilities.**

I. Introduction

People have long anticipated that an intelligent home service robot taking care of their laundry chores, like folding a T-shirt for storage. Despite recent significant advances in clothes manipulation [1], [2], these methods are tailored to specific clothes types and tasks. How can we build a robot for general-purpose clothes manipulation? Clothes are deformable objects with a high-dimensional state space, and different types of clothes exhibit distinct geometric structures. This complexity makes it essential to develop a general state representation.

In this paper, we present semantic keypoints as a general spatial-semantic representation of clothes. Compared with previous keypoint representations, semantic keypoints carry explicit semantic meaning and can be intuitively described using natural language. As a result, semantic keypoints offer a sparse representation and focus on distinct structural features of clothes like sleeves and shoulders, salient for both perception and action. For perception, semantic keypoints are easy to extract and consistent across instances of the same clothes type. For action, semantic keypoints identify where clothes are frequently manipulated.

For general-purpose clothes manipulation (Fig 1), we integrate semantic keypoints with foundation models and develop a CLothes mAnipulation method with Semantic keyPoints (CLASP). CLASP first leverages a vision language model (VLM) and vision foundation models for open-category semantic keypoint extraction. After extracting semantic keypoints, the RGB image marked with semantic keypoints and the language instruction are fed into a VLM for task planning. The VLM determines task completion status and generates a sequence of sub-tasks, each consisting of a basic skill and associated contact points. Before execution, the sub-task plan is verified. After executing each sub-task, CLASP updates its observation and decides whether to replan. This loop will be repeated until the task is completed.

We conduct both simulation and real-world experiments to evaluate CLASP. Simulations show that CLASP outperforms baseline methods in success rate and generalization. Real-world experiments on 15 various clothes show that CLASP performs well across a wide variety of clothes and tasks.

II. Related Work

Learning Deformable Object Manipulation. Although learning methods have made significant progress in a wide range of deformable object manipulation tasks [3]–[5], the generalization remains a significant limitation. Most of the

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

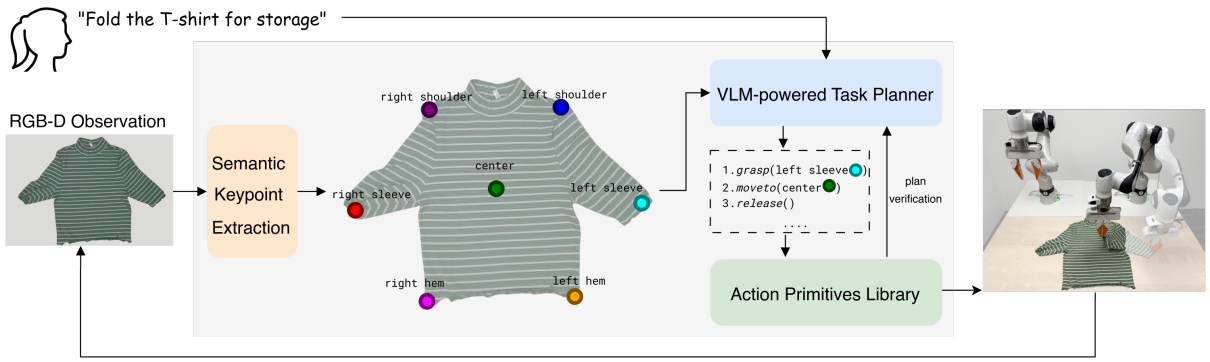


Fig. 2: CLASP overview. Given an RGB-D observation, CLASP extracts semantic keypoints as the state. These keypoints, along with the RGB image and task instruction, are fed to a VLM to generate a sub-task sequence. Once verified, the sub-tasks are executed. After each sub-task, CLASP updates the observation and decides whether to replan. This loop will be repeated until the task is complete.

previous methods learn task-specific policies [6]–[8]. Although some goal-conditioned methods aim at multi-task learning [9], [10], they often struggle to generalize to unseen goals. In this paper, we propose CLASP, a framework that integrates semantic keypoint representations with foundation models. CLASP enables generalization to a wide range of clothes and manipulation tasks.

State representation of deformable objects. To represent deformable objects with high-dimensional state space, physics-based representation [11]–[13], latent representation [14], [15], and keypoint representation [16]–[18] are explored. Among them, Keypoint representation provides a succinct way and enables efficient planning and learning [19]. In this paper, we present semantic keypoints as a general spatial-semantic representation of clothes. Unlike previous keypoint representations, semantic keypoints carry semantic meaning and can be described in natural language, making them sparse and salient for perception and action.

III. Method

In this paper, we introduce CLASP (CLothes mAnipulation method with Semantic keyPoints), a general-purpose clothes manipulation method. The core idea of CLASP is to use semantic keypoints as a general state representation for clothes. Each semantic keypoint includes a language description (e.g., "left sleeve") and a keypoint position.

To leverage semantic keypoints in clothes manipulation, we first build a basic skill library. Instead of manually designing skills, we prompt a large language model (LLM) to discover basic clothes manipulation skills. The LLM's commonsense knowledge ensures these skills are general and sufficient. We then implement the discovered skills as policies parameterized by contact point positions, to form the basic skill library. The basic skill library includes `grasp`, `moveto`, `release`, `rotate`, and `pull`.

Fig. 2 illustrates the overall framework of CLASP. Given RGB-D observations of clothes and language instruction, CLASP generates sequential manipulation actions to complete the task specified by the instruction. CLASP first leverages foundation models for open-category semantic keypoint extraction. Specifically, we propose a two-stage pipeline for semantic keypoint extraction. The first stage leverages a VLM for semantic understanding of clothes and autonomously discovers semantic keypoints on a prototype image for each clothes category. The second stage aims to match semantic keypoints on the prototype image to novel clothes, where vision foundation models ensure spatial precision.

Given the observation image and extracted semantic keypoints, we prompt the VLM to decompose the free-form natural language instruction into sequential subtasks. Each subtask consists of a basic skill (from a predefined skill library) and contact points (selected from semantic keypoints), such as `grasp`(*"left sleeve"*). CLASP then verifies the plan's executability: for each subtask, it invokes a policy from a low-level skill library to generate waypoints based on contact point positions and uses motion planning to produce trajectories. If all subtasks are feasible, we execute them sequentially. Otherwise, the failure reason is used to prompt the VLM for replanning. After executing each subtask, CLASP updates the observation and decides whether to replan. This closed-loop pipeline repeats until the task is completed, as determined by the VLM.

IV. Experiments

A. Simulation Experiments

To evaluate CLASP, we conduct simulation experiments in SoftGym [22]. The tasks include folding, flattening, hanging, and placing. Baseline methods include two general multi-task learning frameworks (CLIPORT [20] and Goal-conditioned Transporter [9]) and two task-specific algorithms (FlingBot [21] and FabricFlowNet [7]). To evaluate the generalization, only half of the tasks are provided through expert demonstrations or examples in the prompt. For each task, we conduct 120 trials with different configurations to calculate the success rate. The two task-specific baseline methods are only evaluated on related tasks.

The experiment results are shown in TABLE I. Overall, CLASP outperforms the two multi-task learning methods on both seen and unseen tasks. The two multi-task learning methods learn task-specific action sequences in an end-to-end manner, and the learned action sequences are not

TABLE I: Simulation Experiment Results. The average success rates (%) on testing tasks. The best performance is in bold.

Method	Folding (seen object)		Flattening (seen object)		Hanging (seen object)		Placing (seen object)	
	Towel	T-shirt	T-shirt	Skirt	Trousers	Towel	Towel	Skirt
CLIPORT [20]	77.5	80.0	32.5	36.7	76.7	83.3	93.3	93.3
Goal-conditioned Transporter [9]	83.3	76.7	26.7	33.3	**100.0**	80.0	66.7	83.3
FlingBot [21]	N/A	N/A	**66.7**	**85.0**	N/A	N/A	N/A	N/A
FabricFlowNet [7]	93.7	**100.0**	N/A	N/A	N/A	N/A	N/A	N/A
CLASP	**100.0**	95.0	65.0	80.0	96.7	**97.5**	**96.7**	**96.7**
Method	Folding (unseen object)		Flattening (unseen object)		Hanging (unseen object)		Placing (unseen object)	
	Trousers	Skirt	Trousers	Towel	T-shirt	Skirt	Trousers	T-shirt
CLIPORT [20]	0.0	0.0	8.3	9.2	76.7	66.7	70.0	73.3
Goal-conditioned Transporter [9]	0.0	0.0	10.0	6.7	36.7	40.0	36.7	60.0
FlingBot [21]	N/A	N/A	29.2	34.2	N/A	N/A	N/A	N/A
FabricFlowNet [7]	0.0	2.5	N/A	N/A	N/A	N/A	N/A	N/A
CLASP	**87.5**	**81.7**	**60.8**	**65.0**	**93.3**	**76.7**	**93.3**	**94.2**

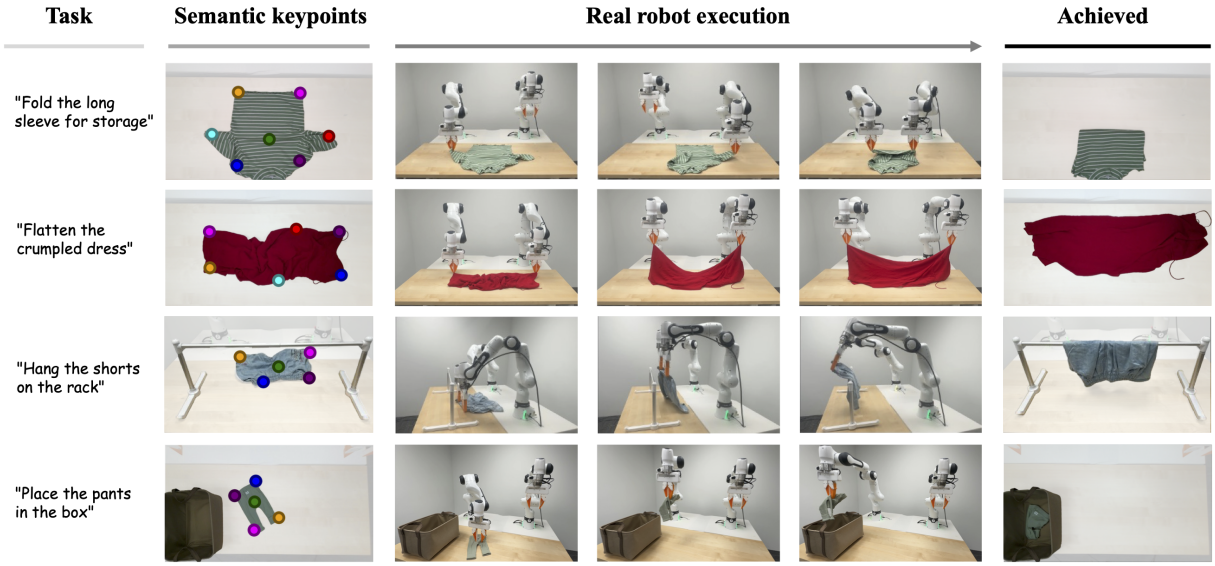


Fig. 3: Qualitative results of real-world experiments. From left to right, the figures display the task description, detected semantic keypoints, the real robot's execution process, and the final achieved state.

transferrable. In contrast, CLASP learns task-agnostic and generalizable language and visual concepts. The commonsense knowledge from VLM allows CLASP to handle unseen clothes manipulation tasks by decomposing them into predefined basic skills. Furthermore, semantic keypoints are task-agnostic and provide effective cues for task planning and action generation in unseen clothes manipulation tasks. Compared to the two task-specific algorithms, CLASP shows comparable performance on seen clothes folding and flattening tasks, demonstrating the effectiveness of the proposed method for clothes manipulation.

B. Real Experiments

In real-world experiments, we evaluate CLASP on 15 clothes (Fig. 1 (a)) across diverse types, sizes, shapes, and materials. Each clothes item is evaluated across four tasks: folding, flattening, hanging, and placing. ClASP achieves an 86% success rate in clothes folding, a 66% success rate in clothes flattening, a 94% success rate in clothes hanging, and a 92% success rate in clothes placing. The success rate is comparable to existing task-specific clothes manipulation algorithms while we test CLASP on broader clothes types and instances. Fig. 3 illustrates some representative examples.

V. Conclusion

In this paper, we present semantic keypoints as a general spatial-semantic representation of clothes. Building on this representation, we propose CLothes mAnipulation with Semantic keyPoints (CLASP). By integrating a general semantic keypoint representation with the capabilities of foundation models, CLASP provides a general-purpose solution for clothes manipulation. Simulation experiments demonstrate that CLASP outperforms baseline methods in terms of success rate and generalization in clothes manipulation. Real-world experiments further validate CLASP's generalization in clothes manipulation. CLASP can be directly applied to a diverse range of clothes types and manipulation tasks.

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] A. Ganapathi, P. Sundaresan, B. Thananjeyan, A. Balakrishna, D. Seita, J. Grannen, M. Hwang, R. Hoque, J. E. Gonzalez, N. Jamali, *et al.*, "Learning dense visual correspondences in simulation to smooth and fold real fabrics," in *IEEE International Conference on Robotics and Automation*, 2021.

[4] 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.

[5] 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.

[6] G. Salhotra, I.-C. A. Liu, M. Dominguez-Kuhne, and G. S. Sukhatme, "Learning deformable object manipulation from expert demonstrations," *IEEE Robotics and Automation Letters*, 2022.

[7] 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.

[8] C. Gao, Z. Li, H. Gao, and F. Chen, "Iterative interactive modeling for knotting plastic bags," in *Conference on Robot Learning*, 2023.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] M. Lippi, P. Poklukar, M. C. Welle, A. Varava, H. Yin, A. Marino, and D. Kragic, "Latent space roadmap for visual action planning of deformable and rigid object manipulation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020.

[15] W. Yan, A. Vangipuram, P. Abbeel, and L. Pinto, "Learning predictive representations for deformable objects using contrastive estimation," in *Conference on Robot Learning*, 2021.

[16] 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.

[17] 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.

[18] B. Zhou, H. Zhou, T. Liang, Q. Yu, S. Zhao, Y. Zeng, J. Lv, S. Luo, Q. Wang, X. Yu, *et al.*, "Clothesnet: An information-rich 3d garment model repository with simulated clothes environment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[19] 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.

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

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

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