Generalizable Clothes Manipulation with Large Language Model

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

***Abstract*— We have seen much recent progress in task-specific clothes manipulation, but generalizable clothes manipulation is still a challenge. Clothes manipulation requires sequential actions, making it challenging to generalize to unseen tasks. Besides, a general clothes state representation method is crucial. In this paper, we adopt language instructions to specify and decompose clothes manipulation tasks, and propose a large language model based hierarchical learning method to enhance generalization. For state representation, we use semantic keypoints to capture the geometry of clothes and outline their manipulation methods. Simulation experiments show that the proposed method outperforms the baseline method in terms of success rate and generalization for clothes manipulation tasks.**

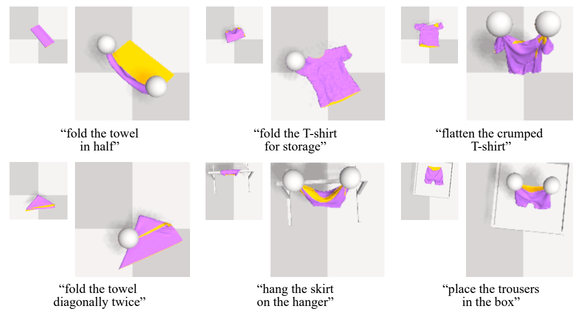


Fig. 1: **Generalizable Clothes Manipulation**. Our proposed method enables generalizable clothes manipulation and is applicable to a wide range of clothes manipulation tasks and object categories.

I. Introduction

People have long anticipated that an intelligent household robot can free them from the tedium of organizing and storing clothes. Toward this goal, the robot should 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). Recently, learning task-specific clothes manipulation skills has been widely investigated [1], [2]. However, these methods often fail to generalize to unseen tasks with new object categories or new requirements such as different folding direction, position, and times. For example, it's difficult to transfer the skill from a T-shirt to a skirt or from folding a towel in half once to folding it twice. However, generalizable clothes manipulation poses two challenges. Clothes manipulation tasks typically require sequential actions, where the action order is crucial for task completion. Thus, generalizing to unseen tasks poses higher requirements for task planning. Besides, clothes are characterized by high dimensionality within their state space [3]. Moreover, the geometric structures of different clothes vary significantly. Thus, an effective and general state representation method is crucial.

For generalizable clothes manipulation, we develop a hierarchical learning method and decompose clothes manipulation into three levels of hierarchy: planning, grounding, and action (Fig. 2). In the planning layer, we adopt language instructions to specify clothes manipulation tasks and use a large language model (LLM) for task planning. Compared to goal images, language instructions provide a more intuitive and flexible way of specifying tasks. Furthermore, large language models can provide commonsense knowledge for task planning and enhance the generalization [4]. Specifically, we prompt the LLM to decompose the given language instruction into sequential sub-tasks. Each sub-task is described by predefined action primitives and contact point descriptions. We use the LLM to condense actions used in its task planning to action primitives and benefit task planning. Contact point descriptions are geometric features of clothes like "left sleeve".

In the grounding layer, we address state representation and visually ground contact point descriptions. Clothes has a predefined structure with significant geometric features, such as sleeves and collars. Semantic keypoints of these geometric features can capture the geometry of clothes and define how they can be manipulated [5]. Thus, we adopt semantic keypoints as the cloth state representation and contact point candidates. To detect effective semantic keypoints, we use a masked auto-encoder [6] to learn a powerful spatiotemporal representation through reconstructing masked image sequences. The pre-trained spatiotemporal representation can handle occlusion since masking is one form of occlusion. We then train the keypoint detector based on the pre-trained representation. After detecting keypoints, visual grounding of contact points is achieved through selecting semantic keypoints based on their semantic meaning. Finally, the action layer will generate trajectories conditioned on action primitives and contact points.

To evaluate the proposed hierarchical learning method, we extend SoftGym [7] benchmark and conduct simulation experiments. Our proposed method outperforms the baseline method in seen and unseen tasks. Hierarchical learning enables the robot to learn transferable language and visual concepts across clothes manipulation tasks, enhancing generalization. In summary, our contributions are as follows:

- We use language instructions for task specification and a large language model for task planning, enabling generalizable clothes manipulation;
- We propose a semantic keypoints based clothes state representation method, leveraging a masked autoencoder to accurately detect keypoints under occlusion.

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

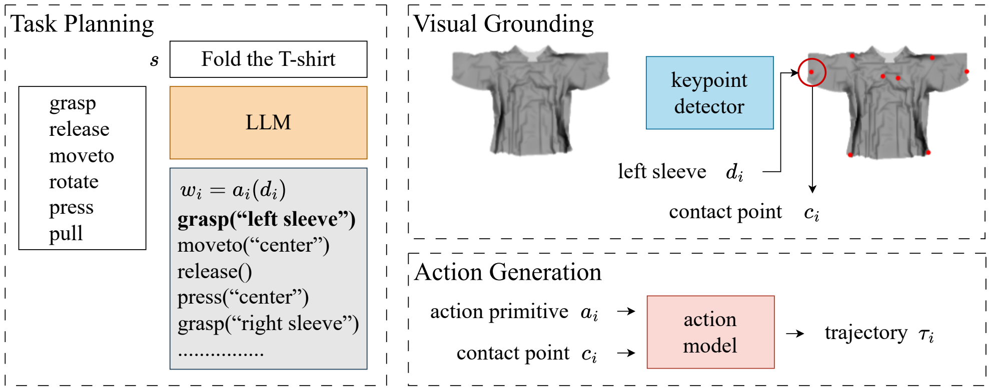


Fig. 2: **Hierarchical learning method**. We decompose the problem of generating action trajectories for clothes manipulation into task planning, visual grounding, and action generation, which enables the robot to learn transferable language and visual concepts across clothes manipulation tasks.

II. Related Work

Learning for Deformable Object Manipulation. Learning methods have been used to equip the robot with task-specific deformable object manipulation abilities such as rope rearrangement [8], [9], cloth folding [1], [10], cloth flattening [11], [12], and bag opening [13], [14]. Some goal-conditioned approaches use goal images to specify different tasks for multi-task learning of deformable object manipulation [15], [16]. However, the task diversity is still limited and it's difficult to generalize to new goals. Unlike previous work, we adopt language instructions to specify and decompose different tasks for generalizable clothes manipulation.

Language-conditioned object manipulation. Language provides an intuitive interface in human-robot interaction and can explicitly capture the transferable concepts between different manipulation tasks. Thus, language-conditioned object manipulation has been widely investigated. Early work focuses on how to make the robot understand language instructions and perform manipulation tasks [17], [18]. Recently, large language models have been employed in language-conditioned manipulation to enhance generalization [19]–[21]. However, previous methods are limited to rigid objects. In this paper, we extend language-conditioned manipulation's application scenarios to deformable objects.

State representation of deformable objects. 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 [22]–[24]. Compared with particles and meshes, keypoints representation has lower dimensionality, leading to more effective policy learning [25]. Keypoints representation is also suitable for clothes, which has a predefined structure with significant geometric features. In this paper, we explore how to detect effective semantic keypoints as the state representation of clothes.

III. Method

In this work, we propose a hierarchical learning method (Fig. 2) that formulates the problem of generating trajectories $\{\tau_i\}$ for clothes manipulation task specified by a given language instruction s into three levels of hierarchy: (1) Task planning – inferring a sequence of sub-task $\{w_i\}$ conditioned on the language instruction s, $w_i = a_i(d_i)$, a_i refers to the action primitive and d_i refers to the language description of the contact point. (2) Visual grounding - for each sub-task w_i, detecting keypoints P_i from observation I_i as the state representation, and grounding contact point c_i conditioned on keypoints P_i and contact point description d_i. (3) Action generation - for each sub-task w_i, generating a trajectory τ_i conditioned on the action primitive a_i and the contact point c_i. We make two assumptions: the clothes manipulation is quasi-static and not long-horizon.

A. Task Planning

LLMs are utilized to enhance robot task planning due to their powerful commonsense knowledge from extensive internet-scale training data. However, previous work is limited to action primitives such as picking and placing, moving, and opening. Such action primitives are not sufficient for generalizable clothes manipulation. Thus, we utilize the LLM with a chain of thought prompting [26] to define action primitives. The LLM is prompted to (1) provide examples of clothes manipulation tasks; (2) decompose these examples into basic actions; (3) summarize the actions used in step (2) and identify action primitives. In this way, we identify action primitives, including *grasp*, *release*, *moveto*, *rotate*, *press*, and *pull*. These action primitives reflect LLM's commonsense knowledge, enhancing LLM's task planning. To generate sub-tasks, we then prompt the LLM with some examples consisting of language instructions paired with desirable sub-tasks sequences.

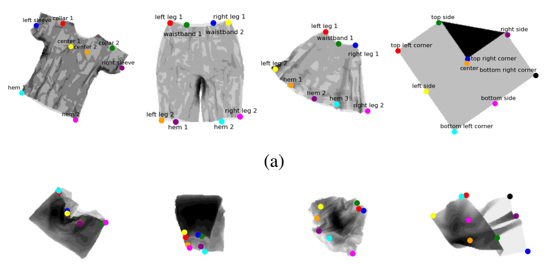
B. Visual Grounding

Upon obtaining a sub-task $w_i = a_i(d_i)$ from task planning, our visual grounding layer will ground the contact point description d_i conditioned on current observation I_i. For sim-to-real transferring, we utilize depth images as observation. Given that clothes has a predefined structure with significant geometric features, such as sleeves and collars, identifying semantic keypoints of these features can effectively capture the clothes' geometry and outline possible manipulation methods. Thus, our visual grounding layer is based on semantic keypoints detection. We first leverage a masked autoencoder as a spatiotemporal learner to establish

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

Method	seen easy tasks			seen medium tasks			seen hard tasks		
	corner folding	half folding	diagonal folding	object hanging	object flattening	object placement	T-shirt folding	trousers folding	all corner folding
CLIPORT [18]	86.7	70.0	76.7	93.3	33.0	**93.3**	80.0	66.7	90.0
Ours	**100.0**	**96.7**	**100.0**	**96.7**	**53.3**	**93.3**	**100.0**	**86.7**	**96.7**
Method	unseen easy tasks			unseen medium tasks			unseen hard tasks		
	corner folding	half folding	diagonal folding	object hanging	object flattening	object placement	skirt folding	half folding twice	diagonal folding twice
CLIPORT [18]	76.7	60.0	70.0	76.7	0.0	80.0	0.0	0.0	0.0
Ours	**100.0**	**96.7**	**96.7**	**83.3**	**50.0**	**93.3**	**83.3**	**93.3**	**96.7**

a powerful latent space to handle occlusion. The core idea is that masking acts as a form of occlusion, and recovering the masked areas requires the model to infer spatial structures of clothes from partial observations. Then, we fine-tune the masked autoencoder with an additional decoder aimed at detecting effective keypoints. As Fig. 3 shows, our keypoint detector can detect effective semantic keypoints of clothes under occlusion. After detecting semantic keypoints from the depth image, the visual grounding of the contact point is achieved by selecting keypoint based on its description.



(a)

(b)

Fig. 3: **Semantic keypoints detection.** (a) Semantic keypoints detection results of flat objects. (b) Semantic keypoints detection results of objects that become occluded after robotic manipulation.

C. Action Generation

After grounding the contact point description d_i to its position c_i, our action model generates an action trajectory τ_i conditioned on the action primitive a_i. The action model is based on manually designed rules.

IV. Experiments

To evaluate the proposed hierarchical learning method on generalizable clothes manipulation, we compared the proposed method with the end-to-end baseline method on a set of clothes manipulation tasks in simulation environment.

A. Simulation Experiment Setup

We choose CLIPORT [18] as the baseline method, which represents the typical end-to-end algorithm for language-conditioned manipulation policy learning. CLIPORT relies on a pre-trained vision-language model.

We extended SoftGym benchmark to 30 common clothes manipulation tasks. These tasks can be mainly divided into: (1) folding clothes in different way (corner folding, half folding, and diagonal folding); (2) folding clothes for storage; (3) flattening crumpled clothes; (4) hanging clothes on a hanger; (5) placing clothes and storing them. The clothes categories include T-shirts, trousers, skirts, and towels. Each category of clothes has over 35 instances with different shapes and sizes. Besides, tasks are categorized by complexity into easy, medium, and hard tasks based on action steps. Easy tasks involve up to 4 steps, middle tasks require 5 to 7 steps, and hard tasks need 8 steps or more. Only half of tasks are seen during training through examples in the prompt or demonstrations. Unseen tasks involves new object categories and new requirements like folding direction, position, and times.

We compare the success rate of different methods on the above tasks. The success metric is the mean particle position error between the clothes states achieved by policy and an oracle demonstrator. We define a task as a success if the mean particle position error is less than the diameter of a particle in the simulation.

B. Simulation Experiment Results

The experiment results are shown in TABLE I. Overall, our method outperforms CLIPORT in seen and unseen tasks, especially when the task complexity increases. CLIPORT can generalize to unseen easy tasks and some unseen medium tasks. The pre-trained vision-language model enable CLIPORT to capture the similarity between different tasks (e.g.“hang the T-shirt” and “hang the skirt”). But it’s difficult to learn action sequences of hard tasks in an end-to-end manner. The learned policies are not generalizable. In contrast, hierarchical learning can learn transferable language and visual concepts across clothes manipulation tasks. LLM can complete the task planning of unseen tasks and decompose unseen tasks to predefined action primitives. Additionally, semantic keypoints are independent of specific tasks, which can be utilized to ground contact points of unseen manipulation tasks.

V. Conclusion

In this paper, we propose a hierarchical learning method for generalizable clothes manipulation, where language instructions and a LLM are used for task specification and planning. To represent the s clothes effectively, we use a masked autoencoder to detect semantic keypoints under occlusion. Semantic keypoints are used to ground contact point of manipulation tasks. Simulation experiment results show that proposed hierarchical learning method outperforms the baseline method in success rate and generalization. Proposed method can generalize to unseen tasks with new object categories or new requirements. For future work, we will explore the generalization on object instances and close-loop task planning of clothes manipulation.

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] J. Matas, S. James, and A. J. Davison, "Sim-to-real reinforcement learning for deformable object manipulation," in *Conference on Robot Learning*, 2018.

[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] 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.

[5] S. Miller, J. Van Den Berg, M. Fritz, T. Darrell, K. Goldberg, and P. Abbeel, "A geometric approach to robotic laundry folding," *The International Journal of Robotics Research*, vol. 31, no. 2, pp. 249–267, 2012.

[6] C. Feichtenhofer, Y. Li, K. He *et al.*, "Masked autoencoders as spatiotemporal learners," *Advances in neural information processing systems*, 2022.

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

[8] 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.

[9] C. Wang, Y. Zhang, X. Zhang, Z. Wu, X. Zhu, S. Jin, T. Tang, and M. Tomizuka, "Offline-online learning of deformation model for cable manipulation with graph neural networks," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5544–5551, 2022.

[10] 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*, vol. 8, no. 2, pp. 760–767, 2022.

[11] 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.

[12] 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.

[13] 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.

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

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

[24] 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.

[25] 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.

[26] 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.