MetaFold: Language-Guided Multi-Category Garment Folding Framework via Trajectory Generation and Foundation Model

Haonan Chen*, Junxiao Li*, Ruihai Wu*, Yiwei Liu, Yiwen Hou, Zhixuan Xu,
Jingxiang Guo, Chongkai Gao, Zhenyu Wei, Shensi Xu, Jiaqi Huang, Lin Shao†

***Abstract*—**
Garment folding is a common yet challenging task in robotic manipulation. The deformability of garments leads to a vast state space and complex dynamics, which complicates precise and fine-grained manipulation. In this paper, we present **MetaFold**, a unified framework that disentangles task planning from action prediction and learns each independently to enhance model generalization. It employs language-guided point cloud trajectory generation for task planning and a low-level foundation model for action prediction. This structure facilitates multi-category learning, enabling the model to adapt flexibly to various user instructions and folding tasks. We also construct a large-scale MetaFold dataset comprising folding point cloud trajectories for a total of 1210 garments across multiple categories, each paired with corresponding language annotations. Extensive experiments demonstrate the superiority of our proposed framework. Supplementary materials are available on our website: https://meta-fold.github.io/.

I. Introduction

Robotic manipulation of deformable objects—such as clothing—remains challenging due to the high-dimensional state space and complex, non-linear fabric dynamics [1], [2]. Inspired by the human separation of high-level planning (brain) from low-level execution (spinal cord), we decompose garment folding into two stages: (1) state planning, by predicting sequences of future point-cloud states during folding; and (2) action execution, by translating these planned states into end-effector motions.

Predicting future states and actions from arbitrary initial states introduces highly complex dynamics, posing significant challenges to the learning process. By contrast, decomposing the task and focusing specifically on predicting garment states during the folding process simplifies the learning process and further facilitates multi-category generalization, benefiting from the strong capabilities of generative models. Although accurately modeling full dynamics is challenging, predicting state transitions for garment folding is comparatively feasible when action prediction is performed separately and conditioned on these transitions.

Accurate prediction of future states in garment manipulation requires an effective representation of garment configurations. Point cloud trajectories fulfill this need by providing a comprehensive spatial model that captures any point in space and encodes temporal changes in object states. Unlike methods that rely on fixed keypoints or skeletal models—often tied to specific garments—our use of point-cloud trajectories preserves complete spatial and temporal information, enabling precise folding across diverse clothing types. [3]–[8]. Additionally, employing generative model to create point cloud trajectories facilitates multi-category learning and generalization within a unified framework, yielding robust performance across diverse garment manipulation tasks. We also leverages language-conditioned trajectory generation, enabling it to adapt dynamically to a wide range of user-specified instructions.

Fig. 1. We present **MetaFold**, a unified framework capable of handling diverse garments and a wide range of language instructions, enabling various clothing folding tasks efficiently.

To bridge planned trajectories with robot control, we integrate a generative trajectory model with the ManiFoundation model [9], which converts point-flow between successive states into contact proposals and motion vectors. This capability allows us to disentangle robot action prediction from the overall manipulation procedure, thereby reducing the complexity of high-level and low-level modules. To enhance model robustness, we propose a closed-loop framework that integrates the point cloud trajectory generation model with the ManiFoundation model for action prediction.

In summary, our main contributions are as follows:

- We propose **MetaFold**, a framework that integrates a language-guided point cloud trajectory generation model with an action prediction foundation model, thereby facilitating multi-category garment folding.
- We developed the point cloud trajectory dataset for fold-

* denotes equal contribution; † denotes the corresponding author

ing garments across multiple categories, accompanied by corresponding language descriptions.
- We conduct extensive experiments demonstrating MetaFold's superior performance in folding accuracy and language generalization.

II. Framework

A. Language-Guided Trajectory Generation

1) Data Generation: Training a point-cloud trajectory generation model requires a dataset. Since existing datasets lack folding trajectories, we use ClothesNetM [10] and the DiffClothAI simulator [11] to simulate deformable garments and generate trajectories. Grasp and target points are chosen heuristically, the grasp point follows a predefined curve, and mesh vertices are extracted and downsampled to form ground-truth point clouds, each annotated with a language description. Our dataset comprises 3376 trajectories over 1210 garments (2664 for training; 712 for testing).

2) Conditional Point Cloud Trajectory Generation: We model garment folding as point-cloud state transitions and implement this "world model" with a CVAE [12] whose encoder and decoder are Transformer blocks. Given an initial point cloud $\mathcal{P} \in \mathbb{R}^{N\times 3}$ and a language instruction $\mathcal{L}$, we extract spatial features $\mathcal{F}_{\mathcal{P}} \in \mathbb{R}^{N\times 128}$ via PointNet++ and encode $\mathcal{L}$ with LLaMA [13] , projecting it to an embedding $\mathcal{F}_{\mathcal{L}} \in \mathbb{R}^{1\times 128}$. These features condition the CVAE to learn a latent distribution z, that, when sampled and combined with spatial features, generates a sequence of future frames $\mathcal{T} = \{\mathcal{P}_i\}_{i=1}^{M} \in \mathbb{R}^{N\times M\times 3}$. We train with ground-truth trajectories; at inference, we sample z and decode trajectories one folding substage at a time (e.g., one sleeve), simplifying the overall process while enabling the model to sequentially address different tasks by precisely guiding each folding stage.. Joint training across all garment categories lets the model capture both shared and garment-specific folding patterns, boosting cross-category generalization.

B. Closed-Loop Manipulation

Our framework employs a closed-loop manipulation strategy that integrates point cloud acquisition, action prediction, and feedback control to achieve robust garment folding.

1) Point Cloud Acquisition: In simulation, we extract mesh vertices; in real settings, we capture RGB-D data, segment with Segment Anything Model 2 (SAM2) [14], and downsample them to a dimension suitable for the trajectory generation model and the ManiFoundation model [9].

2) ManiFoundation Model: We feed point-flow between successive clouds into ManiFoundation to predict contact and direction proposals, decoupling trajectory planning from action prediction for modular training. We fine-tune it on garment folding data with a contact synthesis loss to improve point and force accuracy. To reduce seed variability, we ensemble 160 runs, cluster predictions within ε distance, and pick the modal contact point and force via their mean.

3) Feedback Control: Once ManiFoundation outputs an action, the robot moves the garment, captures the updated point cloud, and feeds it back into the trajectory generator. This closed-loop control lets the system adapt to disturbances and environmental changes.

III. Experiments

We address three questions:

Q1: How does MetaFold perform on garment folding tasks?
Q2: How does disentangling planning and action compare to end-to-end action prediction?
Q3: How well does MetaFold generalize to diverse language instructions?

A. Simulation and Datasets

We use Isaac Sim [16] for accurate fabric dynamics. Evaluation is on our MetaFold test split and zero-shot on CLOTH3D [17]. We test about 500 distinct garments in total.

B. Metrics

- **Rectangularity**: Final area / bounding rectangle area.
- **Area Ratio**: Final area / initial area.
- **Success Rate**: The percentage of trials that satisfy the thresholds for both Rectangularity and Area Ratio.

C. Baselines

- **For Q1**: UniGarmentManip [5] and GPT-Fabric [18].
- **For Q2**: 3D Diffusion Policy (DP3) [19].
- **For Q3**: L.D. (Deng et al. [15]).

D. Results and Analysis

Table I presents the results for garment folding tasks, Table II shows the results for different language guidance (↑ higher is better; ↓ lower is better).

UniGarmentManip's reliance on demonstrations can lead to geometric failures, and GPT-Fabric's LLM–based keypoint selection is often inaccurate (Table I). MetaFold, which is even zero-shot on CLOTH3D dataset, matches or outperforms these baselines by producing neat, compact folds across most metrics, demonstrating strong generalization (Q1). Its modular separation of planning and execution also surpasses end-to-end DP3 [19], better capturing folding dynamics and enhancing robustness (Q2).

We test on seen (training) and unseen instructions. Table II shows MetaFold outperforms L.D. (Deng et al. [15]) under both settings, demonstrating strong language understanding and generalization (Q3).

E. Real-World Experiments

We evaluated MetaFold on a uFactory xArm6 with xArm Gripper and an overhead RealSense D435 (Fig. 1). Using SAM2 [20] to segment the RGB image, we masked and filtered the depth map to obtain the garment's point cloud. This point cloud drives our pipeline to predict folding states and contact directions, which the robot executes sequentially. Thanks to the small sim-to-real gap of point clouds, no additional adaptation was needed. Quantitative outcomes (10 trials per garment) are in Table III.

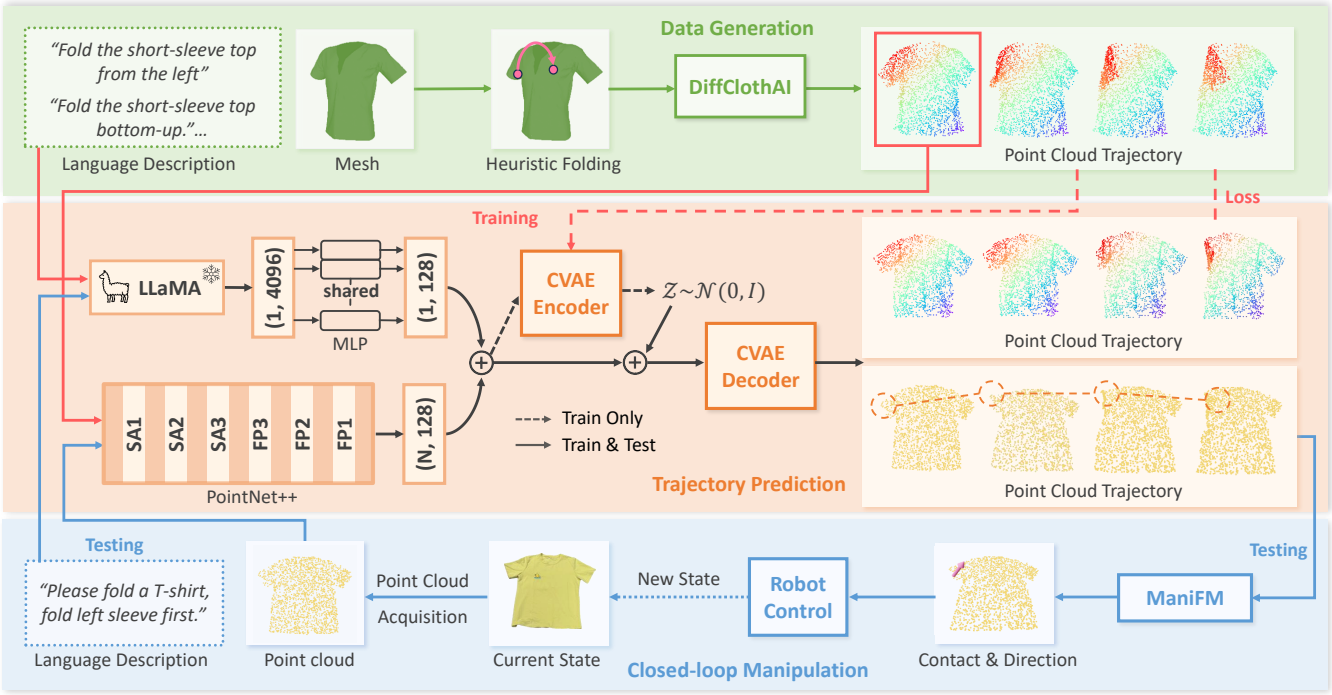


Fig. 2. Overview: The folding trajectory data for clothing is generated using heuristic methods in the DiffClothAI simulation environment, with language descriptions subsequently added **(Green)**. The trajectory generation model takes a point cloud from any given frame and a corresponding language description as inputs to generate the subsequent trajectory **(Orange)**. The generated trajectory is fed into the ManiFoundation model to estimate contact points and force directions, enabling the robot to conduct garment folding actions. This process is then iteratively refined using a feedback loop **(Blue)**.

TABLE I

SIMULATION RESULTS ON GARMENT FOLDING TASKS. UNIG STANDS FOR UNIGARMENTMANIP [5].

		MetaFold Dataset				Cloth3D			
		No-sleeve	Short-sleeve	Long-sleeve	Pants	No-sleeve	Short-sleeve	Long-sleeve	Pants
Rectangularity ↑	UniG	0.85	0.78	**0.88**	0.81	**0.82**	**0.80**	**0.85**	0.83
	DP3	0.85	0.82	0.86	**0.88**	0.80	0.76	0.78	0.79
	GPT-Fabric	0.78	0.78	0.77	0.66	0.81	0.78	0.80	**0.83**
	Ours	**0.87**	**0.83**	0.85	0.86	**0.82**	**0.80**	0.83	**0.83**
Area Ratio ↓	UniG	0.48	0.34	0.34	0.34	**0.47**	0.43	0.34	0.28
	DP3	0.50	0.44	0.39	0.33	**0.47**	**0.33**	0.26	0.28
	GPT-Fabric	0.48	0.45	0.47	0.44	0.54	0.46	0.47	0.50
	Ours	**0.45**	**0.33**	**0.24**	**0.26**	**0.47**	**0.33**	**0.25**	**0.27**
Success Rate ↑	UniG	0.71	0.69	**0.90**	0.77	0.77	0.42	0.71	0.91
	DP3	0.73	0.66	0.37	0.94	0.71	0.70	0.82	0.85
	GPT-Fabric	0.34	0.21	0.03	0.40	0.63	0.22	0.15	0.03
	Ours	**0.88**	**0.86**	**0.90**	**0.97**	**0.79**	**0.86**	**0.97**	**0.97**

TABLE II

DIFFERENT LANGUAGE-GUIDED FOLDING TASKS. L.D. STANDS FOR [15]. "SEEN" AND "UNSEEN" REFER TO THE INPUT INSTRUCTIONS.

		MetaFold Dataset		Cloth3D	
		Seen	Unseen	Seen	Unseen
Rectangularity ↑	L.D.	0.78	0.78	0.81	**0.81**
	Ours	**0.85**	**0.80**	**0.83**	**0.81**
Area Ratio ↓	L.D.	0.36	0.37	0.39	0.40
	Ours	**0.24**	**0.33**	**0.25**	**0.26**
Success Rate ↑	L.D.	0.46	0.46	0.56	0.47
	Ours	**0.90**	**0.63**	**0.97**	**0.93**

TABLE III

REAL WORLD RESULTS FOR OUR FRAMEWORK

	No-sleeve	Short-sleeve	Long-sleeve	Pants
Rectangularity ↑	0.94	0.91	0.87	0.85
Area Ratio ↓	0.45	0.33	0.29	0.24
Success Rate ↑	10/10	8/10	9/10	9/10

IV. CONCLUSION

In this work, we propose a comprehensive framework, **MetaFold**, for garment folding that supports multi-category folding tasks guided by user language instructions. This framework adopts a disentangled structure consisting of a point cloud trajectory generation model and a low-level action prediction model, utilizing closed-loop control for effective garment manipulation. We also constructed a point cloud trajectory dataset for garment folding, encompassing various folding methods across different garment types.

Experimental results demonstrate that our approach achieves state-of-the-art performance in multi-category and language-guided garment folding tasks. We believe that MetaFold represents a significant step forward in applying trajectory generation to deformable object manipulation, marking an important milestone toward advancing spatial intelligence.

REFERENCES

[1] J. Zhu, A. Cherubini, C. Dune, D. Navarro-Alarcon, F. Alambeigi, D. Berenson, *et al.*, "Challenges and outlook in robotic manipulation of deformable objects," *IEEE Robotics & Automation Magazine*, 2022.

[2] F. Gu, Y. Zhou, Z. Wang, S. Jiang, and B. He, "A survey on robotic manipulation of deformable objects: Recent advances, open challenges and new frontiers," *arXiv preprint arXiv:2312.10419*, 2023.

[3] R. Shi, Z. Xue, Y. You, and C. Lu, "Skeleton merger: an unsupervised aligned keypoint detector," in *CVPR*, 2021, pp. 43–52.

[4] A. Ganapathi, P. Sundaresan, B. Thananjeyan, A. Balakrishna, D. Seita, J. Grannen, *et al.*, "Learning dense visual correspondences in simulation to smooth and fold real fabrics," in *ICRA*, 2021.

[5] R. Wu, H. Lu, Y. Wang, Y. Wang, and H. Dong, "Unigarmentmanip: A unified framework for category-level garment manipulation via dense visual correspondence," in *CVPR*, 2024, pp. 16 340–16 350.

[6] J. Hietala, D. Blanco-Mulero, G. Alcan, and V. Kyrki, "Learning visual feedback control for dynamic cloth folding," in *IROS*. IEEE, 2022.

[7] C. He, L. Meng, Z. Sun, J. Wang, and M. Q.-H. Meng, "Fabricfolding: learning efficient fabric folding without expert demonstrations," *Robotica*, vol. 42, no. 4, pp. 1281–1296, 2024.

[8] 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 *ICRA*. IEEE, 2023, pp. 5872–5879.

[9] Z. Xu, C. Gao, Z. Liu, G. Yang, C. Tie, H. Zheng, H. Zhou, *et al.*, "Manifoundation model for general-purpose robotic manipulation of contact synthesis with arbitrary objects and robots," in *IROS*, 2024.

[10] B. Zhou, H. Zhou, T. Liang, Q. Yu, S. Zhao, Y. Zeng, J. Lv, S. Luo, Q. Wang, X. Yu, H. Chen, C. Lu, and L. Shao, "Clothesnet: An information-rich 3d garment model repository with simulated clothes environment," in *ICCV*, October 2023, pp. 20 428–20 438.

[11] X. Yu, S. Zhao, S. Luo, G. Yang, and L. Shao, "Diffclothai: Differentiable cloth simulation with intersection-free frictional contact and differentiable two-way coupling with articulated rigid bodies," in *IROS*, 2023, pp. 400–407.

[12] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *NIPS*, 2015.

[13] H. Touvron, T. Lavril, G. Izacard, X. Martinet, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[14] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[15] Y. Deng, K. Mo, C. Xia, and X. Wang, "Learning language-conditioned deformable object manipulation with graph dynamics," in *ICRA*. IEEE, 2024, pp. 7508–7514.

[16] J. Liang, V. Makoviychuk, A. Handa, N. Chentanez, M. Macklin, and D. Fox, "Gpu-accelerated robotic simulation for distributed reinforcement learning," in *Conference on Robot Learning*. PMLR, 2018.

[17] H. Bertiche, M. Madadi, and S. Escalera, "Cloth3d: clothed 3d humans," in *ECCV*. Springer, 2020, pp. 344–359.

[18] V. Raval, E. Zhao, H. Zhang, S. Nikolaidis, and D. Seita, "Gpt-fabric: Folding and smoothing fabric by leveraging pre-trained foundation models," *arXiv preprint arXiv:2406.09640*, 2024.

[19] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," in *RSS*, 2024.

[20] N. Ravi, V. Gabeur, *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

APPENDIX

A. Problem Formulation

The goal of language-guided garment folding is to generate a sequence of actions $\{\boldsymbol{a}_i\}_{i=1}^n$ to fold the garment into a target point cloud configuration $\mathcal{P}_{goal}$, given the 3D point cloud observation $\mathcal{P} \in \mathbb{R}^{N\times 3}$ with N points and the language guidance $\mathcal{L}$.

As introduced before, we disentangle this task into three sub-tasks: Point cloud trajectory generation, action prediction, and corresponding closed-loop manipulation:

(1) Given the current $\mathcal{P}$ with N points and $\mathcal{L}$, the goal of the point cloud trajectory generation model is to generate the trajectory $\mathcal{T} = \{\mathcal{P}_i\}_{i=1}^M \in \mathbb{R}^{M\times N\times 3}$ that represents the evolution of the point cloud over time, where M is the number of frames.

(2) Given two point clouds $(\mathcal{P}, \mathcal{P}')$, the aim of the ManiFoundation [9] model is to predict the action $\boldsymbol{a} = \{c_i\}_{i=1}^n$, which is defined as a set of contact syntheses. The contact synthesis for end effector i is $c_i = (\boldsymbol{p}, \boldsymbol{s})$, where $\boldsymbol{p} \in \mathbb{R}^3$ is the contact position and $\boldsymbol{s} \in \mathbb{R}^3$ is the corresponding motion direction based on the trajectory. We slice the generated garment point cloud trajectory $\mathcal{T}$ and input the segments into the ManiFoundation model to predict an action $\boldsymbol{a} = \mathcal{M}_{MF}(\mathcal{P}, \mathcal{P}')$, where $\{\mathcal{P}, \mathcal{P}'\} \subseteq \mathcal{T}$.

(3) Afterward, the robot executes the action $\boldsymbol{a}$, manipulating the garment to a new configuration $\mathcal{P}^* = \mathcal{M}_{Robot}(\mathcal{P}, \boldsymbol{a})$.

We perform processes (1), (2), and (3) iteratively until the current point cloud configuration $\mathcal{P}$ matches the desired point cloud configuration $\mathcal{P}_{goal}$.

B. MetaFold Dataset

Our dataset is visualized in Figure 3. The dataset consists of folding point cloud trajectories from a total of 1210 garments and 3376 trajectories, with 2664 trajectories in the training set and 712 in the test set. A seen instruction might be "Fold the short-sleeve top from the left," while an unseen instruction could be "Please fold the garment from the left sleeve."

For details of the dataset composition, please refer to Table V. The dataset is available on Hugging Face at https://huggingface.co/datasets/chenhn02/MetaFold. The dataset and model-generated trajectories are visualized in Figure 5. An interactive visualization can be found on our website at https://meta-fold.github.io/.

Fig. 3. Example garments in our MetaFold Dataset. The garments demonstrate diversities in categories, shapes, and deformations.

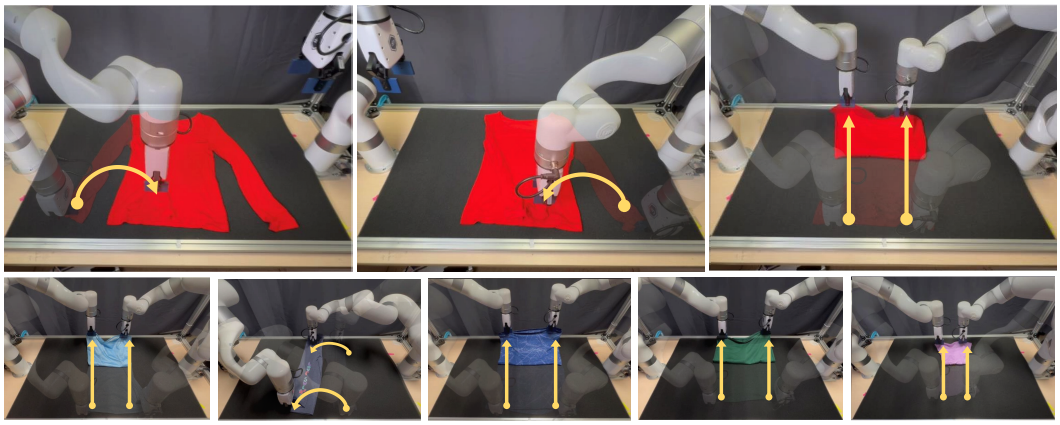


Fig. 4. Real-World Experiments of MetaFold on Diverse Garment Types: Long-Sleeve, Short-Sleeve, No-Sleeve, and Pants.

TABLE IV

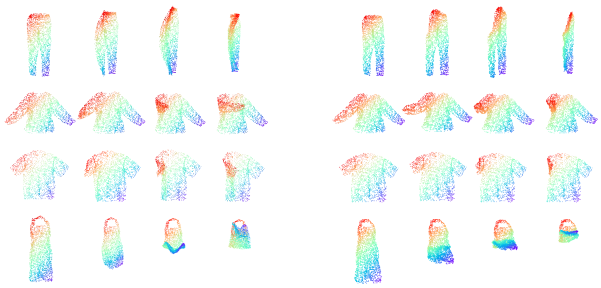
ABLATION STUDIES. ALL ABLATION EXPERIMENTS WERE CONDUCTED ON THE SAME GARMENT TYPE.

Methods	**Ours**	Ours-5frames	Ours-15frames	Ours-NextStep	Ours w/o MF	Ours w/o CL
Rectangularity ↑	**0.83**	0.80	**0.83**	0.79	0.81	0.81
Area Ratio ↓	**0.33**	0.42	0.34	0.35	0.46	0.60
Success Rate ↑	**0.86**	0.51	0.69	0.41	0.27	0.07

TABLE V

DETAILS OF METAFOLD DATASET. S STANDS FOR SLEEVE.

Type	No-S	Short-S	Long-S	Pants	Total
Garments	666	121	146	277	1210
Trajectories	666	726	876	1108	3376



a) Ground truth folding trajectories b) Generated folding trajectories

Fig. 5. Visualization of ground truth and generated trajectories.

C. Language-Guided Folding

Our model supports folding based on languages, even if the folding sequence (Bottom → Left → Right) does not exist during training. Figure 6 illustrates the folding results of our model under different language instructions.

D. Ablation Studies

We compared our framework with several ablated versions to demonstrate the effectiveness of its components:

- **Ours w/o MF**: our method without ManiFoundation model. We randomly select points from the entire set, filter out those with minimal motion, and use grouping to determine contact points. We use the selected point's trajectory direction prediction as the force direction.

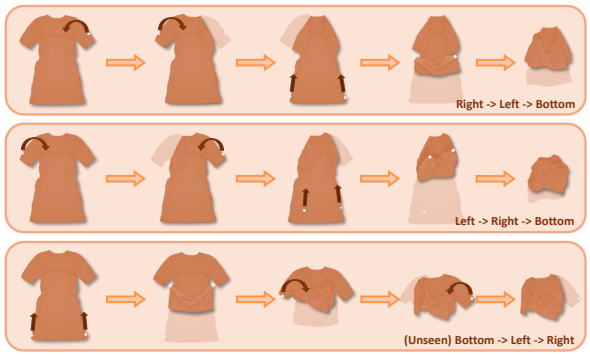


Fig. 6. MetaFold generates different sequences under different instructions.

- **Ours w/o CL**: our method with open-loop control instead of closed-loop control. The execution relies entirely on the initial frame's predicted trajectory.
- **Ours-5frames and Ours-15frames**: represent different granularities of closed-loop execution, performed every 5 frames and every 15 frames, respectively.
- **Ours-NextStep**: predicts only a single step at a time rather than an entire trajectory.

Table IV presents a quantitative comparison with these ablated versions. Experimental results indicate that both the ManiFoundation model and closed-loop control are essential components in garment folding tasks. The results also indicate that our approach achieves optimal performance when closed-loop execution is performed every 10 frames. Predicting the entire trajectory enables the model to generate more effective sequences of actions.

E. Real World Experiment Visualization

Real-world experiments are visualized in Figure 4. Videos of them are available on: https://meta-fold.github.io/.