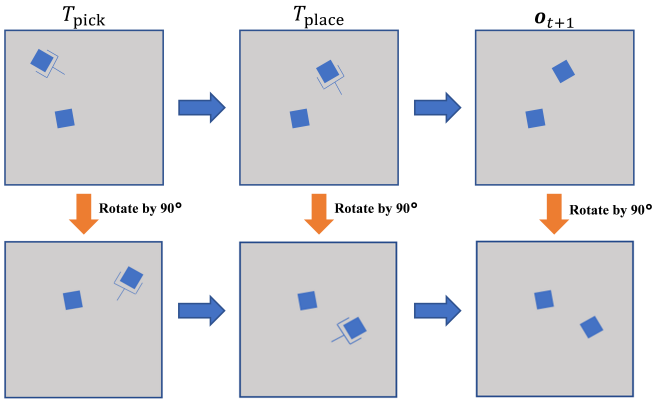

Fig. 7: **Tabletop Rearrangment.** When the current observation and the pick-and-place action are transformed by $g \in SE(2)$, the next-step observation will also be transformed by the same g.

SUPPLEMENTARY MATERIAL

The supplementary material is organized as follows:

- Sec. *A* discusses more details on the equivariance property of the dynamics of tabletop rearrangement.
- Sec. *B* shows detailed results on both training and unseen tasks in simulation. It also contains ablation studies which compare more TVF variants.
- Sec. *C* shows results of simulation experiments with more variable objects.
- Sec. *D* describes more experiment details.

A. SE(2) Equivariance of Dynamics

We assume a pre-defined 2D frame is attached to the infinite tabletop plane. All the coordinates and poses defined below are relative to this frame. Our observation is the orthographic top-down view $\mathbf{o}_t : \mathbb{R}^2 \to \mathbb{R}^4$ of the whole tabletop workspace where $\mathbf{o}_t(u, v)$ gives the observed RGB and height value at position $\mathbf{p} = [u, v]^T \in \mathbb{R}^2$. $g \in SE(2)$ can be parameterized with $g = (R(\theta), \mathbf{q})$ in which $\mathbf{q} = [\Delta u, \Delta v]^T \in \mathbb{R}^2$ represents the translation; $R(\theta)$ represents the rotation:

$$R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{4}$$

The group action of $SE(2)$ on $\mathbf{p} \in \mathbb{R}^2$ and its inverse are defined respectively as:

$$g \diamond \mathbf{p} \doteq R(\theta)\mathbf{p} + \mathbf{q} \tag{5}$$

$$g^{-1} \diamond \mathbf{p} \doteq R(\theta)^{-1}\mathbf{p} - R(\theta)^{-1}\mathbf{q} \tag{6}$$

We define the group action of $SE(2)$ on $\mathbf{o}_t$ as $g \cdot \mathbf{o}_t(\mathbf{p}) \doteq \mathbf{o}_t(g^{-1} \diamond \mathbf{p})$. We denote $\mathbf{x}_t = (\mathbf{o}_t, \mathbf{a}_t)$ where $\mathbf{a}_t = (T_{\text{pick}}, T_{\text{place}}) \in SE(2) \times SE(2)$. The group action of $SE(2)$ on $\mathbf{a}_t$ is defined as $g \odot \mathbf{a}_t \doteq (g \circ T_{\text{pick}}, g \circ T_{\text{place}})$. $\circ$ is the group operation of $SE(2)$ defined as $g_1 \circ g_2 \doteq (R_1 R_2, R_1\mathbf{q}_2 + \mathbf{q}_1)$. We then define the group action on $\mathbf{x}_t$ as $g \bullet \mathbf{x}_t \doteq (g \cdot \mathbf{o}_t, g \odot \mathbf{a}_t)$. The $SE(2)$ equivariance property of the dynamics function $f : \mathbf{x}_t \to \mathbf{o}_{t+1}$ can be written as:

$$f(g \bullet \mathbf{x}_t) = g \cdot f(\mathbf{x}_t) \tag{7}$$

Intuitively, Eq.7 describes the following property of the dynamics of tabletop rearrangement: if the current observation and the picking and placing poses are transformed by $g \in SE(2)$, the next-step observation should also be transformed by g (Fig.7). Our visual foresight (VF) model achieves translational equivariance by using a fully convolutional network (FCN) as the network architecture. We leave the extension to $SE(2)$ equivariance as future work. A promising direction is to represent the input (*i.e.*, the observation and action) in a way such that it is compatible with an $SE(2)$ equivariant network architecture.

B. Detailed Simulation Experiment Results

In Tab. VIII & IX, we show full testing results on 6 training tasks and 8 unseen tasks. We show both the success rate and rate of progress for each task. For the rate of progress, partial credit is also given to trials which are partially completed. The rate of progress is defined as $\frac{\#\text{of blocks in target poses}}{\#\text{of blocks}}$.

TVF variants outperform GCTN in general, even with only one-step foresight (TVF-Small). The advantage of TVF variants over GCTN is more substantial on unseen tasks. With the increase of demo number per task, the success rates of all methods grow in general. A substantial performance improvement is observed when the demo number per task increases from 1 to 10. When the demo number increases further, the improvement is modest. Given 100 and/or 1000 demos per task, the success rates for some tasks even decrease with the increase of demo number. Similar results are also reported in [9]. With the increase of tree depth, TVF-Large outperforms TVF-Small in general.

Similar conclusions can also be drawn from Tab. X-XII where we show more results on more TVF variants. In these tables, we name each method with three letters "K", "M", and "G", which represent the number of clusters in K-Means Clustering, the number of steps expanded by the multi-modal action proposal module, and the number of steps expanded by GCTN. TVF-Small and -Large correspond to TVF-K2-M1-G0 and TVF-K3-M3-G0, respectively. One additional observation from these tables is that the performance does not always improve with the increase of depth. This counter-intuitive result will be explained in the next paragraph. Another conclusion is that using more clusters for the action proposal improves performance in general.

The reason why in some cases the performance does not improve together with the increase of tree depth is threefold:

1) If all the proposed actions are wrong at a given depth, the performance will not improve with the increase of tree depth.
2) If the task can be finished within very a few steps, larger tree depth will not improve results.
3) If the prediction of the dynamics model becomes unreliable as the tree grows, larger tree depth may deteriorate the performance.

Therefore, to achieve better results with larger tree depth, the action proposal module and the dynamics model should improve simultaneously. One future direction is to improve the generalization capability.

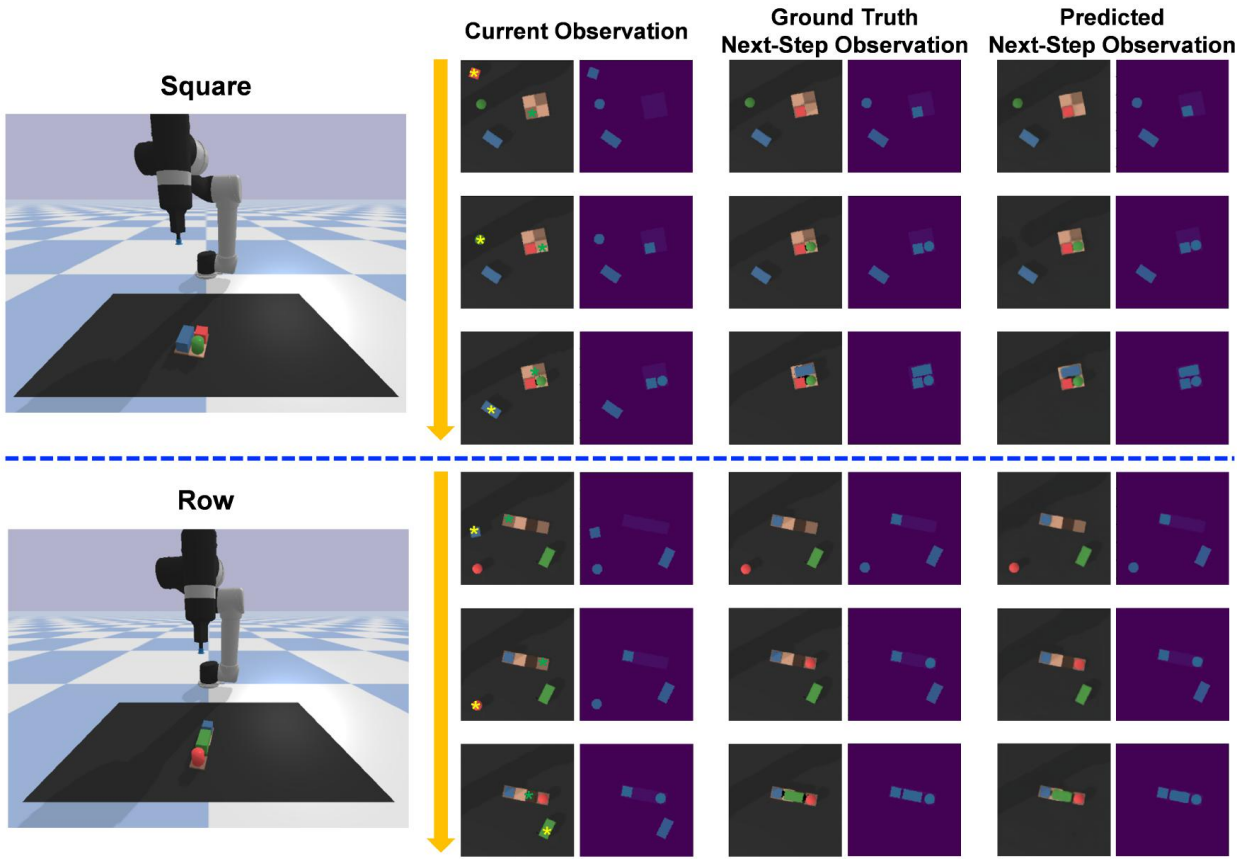

Fig. 8: **VF Model Experiments on Variable Object Shapes and Colors.** We show a rollout on the test data of each of the two tasks, Square and Row. Current Observation shows the RGB image and heightmap of the current step. The yellow and green stars indicate the picking position and placing position, respectively. Ground Truth Next-Step Observation shows the ground truth next-step RGB image and heightmap. Predicted Next-Step Observation shows the predicted next-step RGB image and heightmap of the next step. The actions are expert actions.

C. Experiments on Variable Objects

We perform simulation experiments with more variable objects. In particular, we experiment on two rearrangement tasks, Square and Row, containing a block, a cuboid, and a cylinder (Fig. 8). The colors of the objects in the two tasks also vary. In Square, the block, cuboid, and cylinder are painted red, blue, and green, respectively. In Row, the block, cuboid, and cylinder are painted blue, green, and red, respectively. 10 demos per task are provided as the training data (20 demos in total). Similar to the simulation experiments in V-D, two random actions, which pick an object on the tabletop and place it at a collision-free pose, are also included in the collection of each expert demonstration. We use this data to train the VF model and the GCTN for multi-modal action proposal. Both random actions and oracle actions are used for training the VF model; only oracle actions are used for training the GCTN for the multi-modal action proposal.

We first evaluate our VF model on the test data of these two tasks and compare with the baseline method Latent Dynamics. Qualitative results of our VF model are shown in Fig. 8. Quantitative results are shown in Tab. V. Our VF model is able to retain the data efficiency and performance with variable object shapes and colors. It outperforms Latent Dynamics by a large margin. From Fig. 8, our VF model is able to predict accurate next-step RGB images and heightmaps given current observation and the pick-and-place action. We have also observed that the color prediction of our VF model is worse than that tested on data that contains only red blocks in Sec. V-F (L1 loss of 0.0296 compared to 0.0242 in Tab. IV; lower is better). This is expected because there are more colors in this experiment, making color prediction more challenging. The height prediction is better than that tested on data with only red blocks (L1 loss of 0.0100 compared to 0.0136 in Tab. IV). This is also within expectation because there are only two tasks in this experiment while there are six tasks in Sec. V-D.

We also test the TVF-Small variant on the test data of these two tasks. We compare with GCTN. Results are shown in Tab. VI. TVF-Small outperforms GCTN on both tasks. In both tasks, the success rates of TVF-Small are higher than 80%. Our task planning method is able to retain the data efficiency and performance on tasks with variable objects and colors.

D. More Experiment Details

To collect training data in real robot experiments (Sec. V-E), we implement an efficient way for a human expert to teleoperate the robot to pick and place blocks. See Fig. 9 and its caption for a detailed description of data collection

TABLE V: **Visual Foresight Prediction Results of Experiments on Variable Object Shapes and Colors.** The table shows the visual foresight prediction results of testing our VF model and Latent Dynamics on the test data with variable objects and colors. Both methods are trained with 10 demos per training task (20 demos per task). The table shows the L1 loss of the RGB channels and the height channel between the predicted observation and the ground truth observation. The images are normalized. The actions are expert actions. Lower is better.

Method	Color	Height
Latent Dynamics	0.1082	0.0935
Ours	**0.0296**	**0.0100**

TABLE VI: **Success Rates of Experiments on Variable Object Shapes and Colors.** The table shows the success rates (%) of testing TVF-Small and GCTN on two tasks with variable objects and colors. Each task is tested on 20 rollouts.

Method	Square	Row
GCTN	81.7	68.3
TVF-Small (Ours)	**85.0**	**81.7**

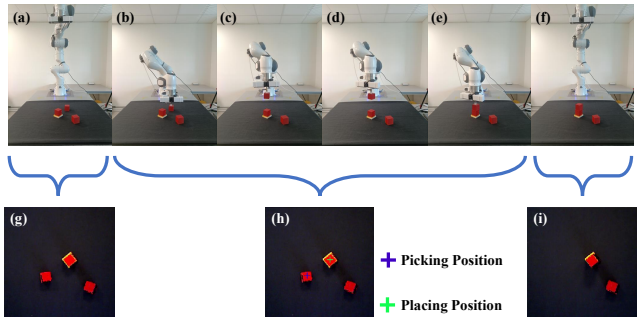

Fig. 9: **Real Robot Data Collection of a Step.** (a) In each step of a rollout, the robot arm moves to the top of the workspace and captures the top-down view RGB image and heightmap. (g) shows the captured top-down RGB image. The top-down image will then show up on the computer and a human expert clicks on two points on the image to specify the picking and placing positions, respectively. (h) shows the clicked points. (b) The robot then moves to the picking position and picks up the block. (c) It then moves towards the placing position. (d) Before placing, the human expert manually specifies the placing rotation angle. (e) Finally, the robot places the block down at the placing position. (f) The robot moves to capture the top-down view and start a new step. If the task is completed, this view will be saved as the goal. (i) shows the top-down RGB image of the observation in (f). The process is repeated until the task is completed.

of a step. Fig. 10 shows the data collection of a rollout. The top-down RGB image and heightmap, the picking and placing positions, and the placing rotation angle of each step are collected during a rollout. The top-down RGB image and heightmap at the end of the rollout are also collected as the goal. To increase data variability, two random actions are also collected in each rollout similar to the simulation experiments in Sec. V-D. The human expert teleoperate the robot to pick a block and place it at a random pose which is collision-free. We perform a background filtering for both RGB images and heighmaps during data collection and testing. Specifically, we convert RGB to HSV and set thresholds on the height and V value. For each filtered pixel, we set the RGB and height value as zero. We find that GCTN is not able to learn well on real data without the background

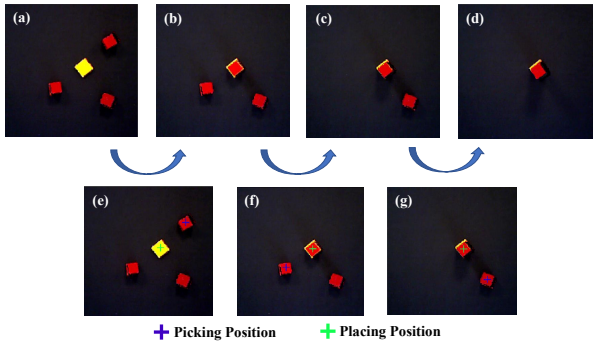

Fig. 10: **Real Robot Data Collection of a Rollout.** (a), (b), (c), and (d) show the top-down RGB images from the initial to the end of a rollout. The task is Tower. (e), (f), and (g) show the picking and placing positions of the three steps specified by the human expert. Random actions are not shown.

TABLE VII: **Hyperparameters**

Hyperparameter	Value
Learning Rate (VF)	1×10^{-4}
Minibatch Size (VF)	1
Training Steps (VF)	6×10^{4}
Learning Rate (Latent Dynamics)	1×10^{-4}
Minibatch Size (Latent Dynamics)	1
Training Steps (Latent Dynamics)	6×10^{4}
Tree Value Coefficient C (TVF)	1
Discount Factor γ (TVF)	0.99
K-Means Clustering Threshold Coefficient α (TVF)	0.01
Top N number in K-Means Clustering N (TVF)	100
Number of Rotation Bin for GCTN R (GCTN)	36

filtering.

Tab. VII shows the hyperparameters we use for training our VF model, GCTN, and Latent Dynamics in the paper. More details can be found in our project website: https://chirikjianlab.github.io/tvf/

TABLE VIII: **Simulation Experiment Results on Training Tasks.** We show the average success rate (%) / rate of progress (%) on the test data of each training task v.s. # of demonstrations (1, 10, 100, or 1000) per task in the training data. Higher is better.

	Row				Square			
Method	1	10	100	1000	1	10	100	1000
GCTN	8.3/35.0	98.3/99.4	**95.0/98.3**	**100.0/100.0**	0.0/34.2	93.3/96.7	65.0/84.2	93.3/96.7
TVF-Small	11.7/**42.2**	**100.0/100.0**	**95.0/98.3**	**100.0/100.0**	1.7/37.1	90.0/95.0	80.0/90.8	98.3/99.2
TVF-Large	**15.0/42.2**	**100.0/100.0**	**95.0/98.3**	**100.0/100.0**	**3.3/40.0**	**100.0/100.0**	**90.0/97.1**	**100.0/100.0**
	T-shape				Tower			
Method	1	10	100	1000	1	10	100	1000
GCTN	1.7/34.2	80.0/93.3	95.0/98.7	95.0/**98.7**	3.3/32.8	**100.0/100.0**	98.3/98.3	**100.0/100.0**
TVF-Small	**3.3**/36.2	90.0/95.8	**96.7/99.2**	95.0/97.5	**5.0/42.8**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**
TVF-Large	1.7/**37.1**	**96.7/98.7**	**96.7/99.2**	**96.7/98.7**	**5.0/42.8**	**100.0/100.0**	**100.0/100.0**	98.3/98.9
	Pyramid				Palace			
Method	1	10	100	1000	1	10	100	1000
GCTN	0.0/31.4	73.3/93.1	83.3/**96.7**	**81.7/95.6**	0.0/32.4	61.7/88.8	78.3/95.2	**85.0/96.4**
TVF-Small	**1.7/34.7**	75.0/**93.3**	**85.0**/96.1	61.7/88.1	0.0/**36.9**	**75.0**/91.9	80.0/95.7	80.0/96.0
TVF-Large	0.0/34.2	**80.0**/92.8	81.7/95.6	66.7/89.2	0.0/33.8	71.7/**92.6**	**85.0/96.9**	83.3/95.5

TABLE IX: **Simulation Experiment Results on Unseen Tasks.** We show the average success rate (%) / rate of progress (%) on the test data of each unseen task v.s. # of demonstrations (1, 10, 100, or 1000) per task in the training data. Higher is better.

	Plane Square				Plane T			
Method	1	10	100	1000	1	10	100	1000
GCTN	1.7/43.8	86.7/96.3	95.0/98.7	96.7/98.7	5.0/39.4	78.3/92.8	93.3/97.8	90.0/96.7
TVF-Small	3.3/**45.0**	98.3/99.6	**96.7/99.2**	**100.0/100.0**	3.3/43.3	**90.0/96.7**	**100.0/100.0**	**98.3/99.4**
TVF-Large	**5.0/45.0**	**100.0/100.0**	**96.7/99.2**	98.3/99.2	**15.0/46.1**	86.7/95.6	**100.0/100.0**	95.0/98.3
	Stair 2				Twin Tower			
Method	1	10	100	1000	1	10	100	1000
GCTN	3.3/38.3	85.0/95.0	46.7/82.2	68.3/89.4	0.0/25.8	88.3/94.7	55.0/71.9	85.0/95.8
TVF-Small	**6.7/43.9**	**98.3/99.4**	71.7/90.6	90.0/96.7	0.0/**34.2**	**98.3/98.9**	**85.0**/90.0	**93.3/97.5**
TVF-Large	3.3/40.0	96.7/97.8	**95.0/97.8**	**100.0/100.0**	0.0/32.5	96.7/98.1	**85.0/91.7**	91.7/96.1
	Stair 3				Building			
Method	1	10	100	1000	1	10	100	1000
GCTN	0.0/30.6	45.0/86.4	23.3/67.5	16.7/76.9	0.0/26.3	5.0/55.3	0.0/57.0	3.3/54.3
TVF-Small	0.0/**38.6**	63.3/91.1	33.3/75.3	46.7/85.0	0.0/**30.3**	8.3/**58.7**	**10.0/66.3**	11.7/64.7
TVF-Large	0.0/32.8	**81.7/94.7**	**56.7/86.9**	**90.0/97.2**	0.0/26.3	**13.3**/58.0	6.7/60.0	**25.0/68.3**
	Pallet				Rectangle			
Method	1	10	100	1000	1	10	100	1000
GCTN	0.0/31.5	23.3/82.5	51.7/78.5	31.7/84.0	0.0/31.1	31.7/84.7	26.7/68.3	41.7/79.4
TVF-Small	0.0/**34.2**	60.0/91.5	61.7/83.5	65.0/94.0	0.0/**35.3**	55.0/88.1	40.0/77.2	75.0/89.7
TVF-Large	0.0/32.1	**75.0/94.4**	**70.0/91.7**	**90.0/97.5**	0.0/30.8	**78.3/94.2**	**63.3/86.4**	**95.0/98.6**

TABLE X: **Ablation Study (10 Demos).** We show the average success rate (%) on the test data of unseen tasks. The number of demonstrations per task in the training data is 10. Higher is better.

Method	Plane Square	Plane T	Stair 2	Twin Tower	Stair 3	Building	Pallet	Rectangle
TVF-K2-M1-G0	98.3	90.0	98.3	98.3	63.3	8.3	60.0	55.0
TVF-K2-M2-G0	100.0	86.7	93.3	98.3	70.0	3.3	65.0	55.0
TVF-K2-M3-G0	100.0	85.0	93.3	95.0	70.0	8.3	56.7	55.0
TVF-K2-M4-G0	100.0	85.0	93.3	95.0	63.3	6.7	68.3	61.7
TVF-K2-M4-G1	100.0	85.0	93.3	98.3	65.0	8.3	66.7	58.3
TVF-K3-M1-G0	100.0	88.3	100.0	98.3	68.3	3.3	76.7	76.7
TVF-K3-M2-G0	100.0	88.3	96.7	90.0	76.7	3.3	66.7	76.7
TVF-K3-M3-G0	100.0	86.7	96.7	96.7	81.7	13.3	75.0	78.3
TVF-K3-M4-G0	100.0	86.7	96.7	95.0	71.7	10.0	70.0	86.7
TVF-K3-M4-G1	100.0	86.7	96.7	95.0	76.7	13.3	68.3	86.7

TABLE XI: **Ablation Study (100 Demos).** We show the average success rate (%) on the test data of unseen tasks. The number of demonstrations per task in the training data is 100. Higher is better.

Method	Plane Square	Plane T	Stair 2	Twin Tower	Stair 3	Building	Pallet	Rectangle
TVF-K2-M1-G0	96.7	100.0	71.7	85.0	33.3	10.0	61.7	40.0
TVF-K2-M2-G0	96.7	100.0	75.0	80.0	38.3	1.7	60.0	43.3
TVF-K2-M3-G0	96.7	100.0	85.0	80.0	43.3	1.7	60.0	51.7
TVF-K2-M4-G0	96.7	100.0	85.0	83.3	41.7	8.3	63.3	51.7
TVF-K2-M4-G1	96.7	100.0	85.0	88.3	43.3	5.0	58.3	51.7
TVF-K3-M1-G0	95.0	100.0	80.0	88.3	51.7	10.0	63.3	53.3
TVF-K3-M2-G0	96.7	100.0	91.7	95.0	55.0	1.7	68.3	60.0
TVF-K3-M3-G0	96.7	100.0	95.0	85.0	56.7	6.7	70.0	63.3
TVF-K3-M4-G0	96.7	100.0	93.3	88.3	53.3	10.0	63.3	68.3
TVF-K3-M4-G1	96.7	100.0	93.3	90.0	51.7	11.7	63.3	66.7

TABLE XII: **Ablation Study (1000 Demos).** We show the average success rate (%) on the test data of unseen tasks. The number of demonstrations per task in the training data is 1000. Higher is better.

Method	Plane Square	Plane T	Stair 2	Twin Tower	Stair 3	Building	Pallet	Rectangle
TVF-K2-M1-G0	100.0	98.3	90.0	93.3	46.7	11.7	65.0	75.0
TVF-K2-M2-G0	98.3	98.3	90.0	96.7	58.3	5.0	76.7	80.0
TVF-K2-M3-G0	100.0	98.3	91.7	98.3	50.0	10.0	75.0	88.3
TVF-K2-M4-G0	100.0	98.3	91.7	98.3	61.7	26.7	71.7	88.3
TVF-K2-M4-G1	100.0	98.3	91.7	96.7	56.7	33.3	80.0	95.0
TVF-K3-M1-G0	100.0	96.7	100.0	98.3	78.3	15.0	88.3	83.3
TVF-K3-M2-G0	98.3	95.0	100.0	98.3	85.0	13.3	93.3	93.3
TVF-K3-M3-G0	98.3	95.0	100.0	91.7	90.0	25.0	90.0	95.0
TVF-K3-M4-G0	100.0	95.0	98.3	95.0	91.7	36.7	88.3	86.7
TVF-K3-M4-G1	100.0	95.0	98.3	98.3	83.3	40.0	86.7	86.7