SUPPLEMENTARY MATERIAL

The supplementary material is organized as follows:

- Section *A* shows testing results on both training and unseen tasks in simulation. It also contains ablation studies which compare more TVF variants trained with different demo numbers.
- Section *B* discusses more details on the equivariance property of the dynamics of tabletop rearrangement.
- Section *C* describes more experiment details.

A. Simulation Results

In Tab. VI & VII, 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. VIII-X 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 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 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. Indeed, on training tasks, methods with larger tree depths and larger demo numbers achieve better results (Tab. VI). However, this pattern is not as obvious for unseen tasks (Tab. VII-X). We believe this is due to the well-known out-of-distribution problem: using more training data on training tasks will not improve the performance on unseen tasks. To further improve the performance on unseen tasks, one future direction is to improve the generalization capability.

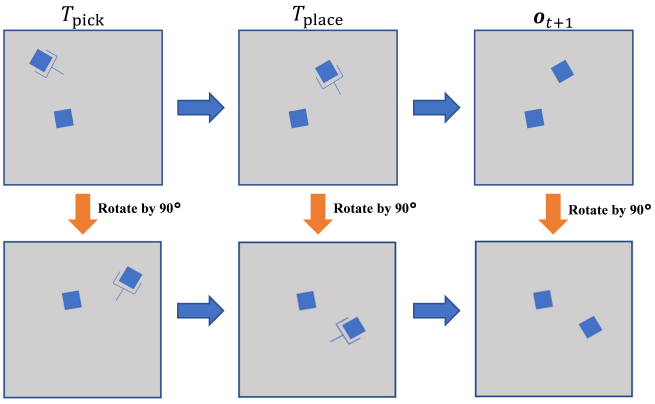

Fig. 7: **Tabletop Rearrangment.** When the current observation and the pick-and-place action are transformed by $g \in SE(2)$, the next-step observation will also be transformed by the same g.

B. 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 $g_1 \circ g_2 \doteq (R_1R_2, R_1\mathbf{q}_2 + \mathbf{q}_1)$. 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: when 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

TABLE V: **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

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.

C. More Experiment Details

In real robot experiments, we pre-process the top-down observation reconstructed from the captured RGB-D image by filtering out the tabletop background. Specifically, we convert RGB to HSV and threshold each pixel by the height and V value. For each filtered pixel, we set the RGB and height value as zero. We observe in the experiment that GCTN is not able to learn well without pre-processing. Tab. V shows the hyperparameters we used 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 VI: **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 VII: **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 VIII: **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 IX: **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 X: **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