Models	Open Drawer	Slide Block	Sweep to Dustpan	Meat off Grill	Turn Tap	Put in Drawer	Close Jar	Drag Stick	Stack Blocks	Screw Bulb
PerAct	80	72	56	84	80	68	60	68	36	24
PerAct(RVT)	88.0 ± 5.7	74.0 ± 13.0	52.0 ± 0.0	70.4 ± 2.0	88.0 ± 4.4	51.2 ± 4.7	55.2 ± 4.7	89.6 ±	26.4 ± 3.2	17.6 ± 2
RVT	71.2 ± 6.9	81.6 ± 5.4	72.0 ± 0.0	88.0 ± 2.5	93.6 ± 4.1	88.0 ± 5.7	52 ± 2.5	**99.2** ± 1.6	28.8 ± 3.9	48.0 ± 5.7
Act3D	93	93	92	**94**	94	**90**	**92**	92	12	47
Ours	**94.4** ± 3.6	**97.6** ± 4.4	**92.8** ± 1.8	90.4 ± 2.2	**96.8** ± 3.3	83.2 ± 1.8	88 ± 2.8	84.8 ± 3.5	**48** ± 0.0	**66.4** ± 2.1

Models	Put in Safe	Place Wine	Put in Cupboard	Sort Shape	Push Buttons	Insert Peg	Stack Cups	Place Cups	Avg. Success	Inf. Speed
PerAct	44	12	16	20	48	0	0	0	42.7	-
PerAct(RVT)	84.0 ± 3.6	44.8 ± 7.8	28.0 ± 4.	16.8 ± 4.7	92.8 ± 3.0	5.6 ± 4.	2.4 ± 2	2.4 ± 3.2	49.4	4.9
RVT	91.2 ± 3.0	**91.0** ± 5.2	49.6 ± 3.2	36.0 ± 2.5	**100.0** ± 0.0	11.2 ± 3.0	**26.4** ± 8.2	4.0 ± 2.5	62.9	11.6
Act3D	**95**	80	**51**	8	99	**27**	9	3	65.1	-
Ours	67.2 ± 3.3	79.6 ± 0.0	32.8 ± 3.3	**48.0** ± 0.0	89.6 ± 1.6	21.6 ± 3.6	18.2 ± 1.8	**13.2** ± 1.8	**67.4**	18.3

Table 1: **Multi-Task Performance on RLBench-100.** Due to the flexibility of our model, it can be easily extended to multi-view inputs. Despite the lack of depth information and targeted design for 3D manipulation, our average success rate still achieves SOTA on RLBench-100.

Dataset ABC→D	Tasks completed in a row					
Method	1	2	3	4	5	Avg.Len.
MT-R3M	0.529	0.234	0.105	0.043	0.018	0.93
GR-1	85.4	71.2	59.6	49.7	40.1	3.06
VidMan(Ego4D)	88.7	77.5	63.8	54.1	45.3	3.29
VidMan(OXE)	95.9	81.6	73.5	61.2	55.1	3.672.3

Table 2: **CALVIN Benchmark Results.** In the video prediction pre-training stage, using the same non-optimal pre-training dataset EGO4D, our model's average length is 0.23 higher than GR-1. When using the expert-annotated OXE data for pre-training, our model achieved an even greater improvement, reaching 3.67.

Google Robot	Pick Coke Can	Pick Object	Move Near	Open Drawer	Close Drawer	Place in Closed Drawer	Avg. Success
RT-1-X	0	0	0	0	0.12	0	0.02
Octo-small	0.24	0.04	0.04	0.04	0.32	0	0.11
Octo-base	0.04	0.08	0	0	0.44	0	0.09
Ours	0.32	0.12	0.04	0.12	0.4	0.08	0.18

Widowx	Spoon on Towel	Carrot on Plate	Stack Cube	Put Eggplant in Basket	Avg. Success
RT-1-X	0	0	0	0.04	0.01
Octo-small	0.48	0.08	0.04	0.6	0.30
Octo-base	0.12	0.04	0	0.32	0.12
ours	0.56	0.2	0	0.6	0.34

Table 3: **Evaluation VidMan trained on the OXE dataset in the SIMPLER environments.** Our method performs better compared to real-world robot manipulation policies (e.g., RT-1-X, Octo) trained on the same dataset in simulation under common setups (e.g., Google Robot, WidowX).

Figure 1: **Video prediction results on OXE.** To convey the dynamics of predicted frames effectively, We reduced the sampling frequency to one-third of that in the manuscript.