Table 1: ID and OOD test grid accuracy with SEM (Standard Error of the Mean) as error bars for the *EnvGen* experiments part of settings G1-G5 and for the Grid-ViT and LLaDA models.

		Grid-ViT		LLaDA	
Setting	**Experiment**	**ID**	**OOD**	**ID**	**OOD**
G1	experiment-1	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	99.9 ± 0.1
	experiment-2	99.7 ± 0.0	92.7 ± 1.3	99.0 ± 0.3	83.5 ± 3.6
	experiment-3	99.8 ± 0.1	95.3 ± 0.4	99.1 ± 0.1	84.2 ± 2.6
	experiment-4	99.2 ± 0.1	88.5 ± 2.5	99.6 ± 0.0	88.9 ± 1.4
	experiment-5	97.8 ± 0.5	73.9 ± 4.9	99.8 ± 0.1	94.1 ± 1.5
G2	experiment-1	100.0 ± 0.0	90.5 ± 4.9	100.0 ± 0.0	60.6 ± 6.6
	experiment-2	97.1 ± 1.8	68.5 ± 3.6	95.5 ± 1.3	45.3 ± 3.4
	experiment-3	97.9 ± 2.1	90.6 ± 5.8	97.7 ± 1.8	47.8 ± 16.8
	experiment-4	99.5 ± 0.2	94.2 ± 1.1	99.7 ± 0.1	89.7 ± 1.2
	experiment-5	95.6 ± 1.1	41.4 ± 6.2	99.6 ± 0.2	68.7 ± 7.8
G3	experiment-1	100.0 ± 0.0	99.7 ± 0.3	100.0 ± 0.0	99.1 ± 0.5
	experiment-2	33.4 ± 33.2	0.0 ± 0.0	18.8 ± 18.5	0.0 ± 0.0
	experiment-3	97.5 ± 0.3	0.1 ± 0.0	96.8 ± 0.6	0.1 ± 0.1
	experiment-4	98.2 ± 1.0	0.0 ± 0.0	97.2 ± 1.3	0.0 ± 0.0
	experiment-5	95.7 ± 1.4	33.1 ± 2.4	99.2 ± 0.5	34.7 ± 0.8
G4	experiment-1	100.0 ± 0.0	79.4 ± 9.2	100.0 ± 0.0	99.6 ± 0.3
	experiment-2	79.6 ± 5.2	0.0 ± 0.0	28.3 ± 28.2	0.0 ± 0.0
	experiment-3	61.8 ± 30.9	0.0 ± 0.0	88.5 ± 3.6	0.9 ± 0.2
	experiment-4	97.0 ± 1.0	15.2 ± 6.3	91.6 ± 2.2	24.7 ± 2.9
	experiment-5	96.0 ± 0.8	0.1 ± 0.1	98.3 ± 0.7	34.7 ± 6.7
G5	experiment-1	100.0 ± 0.0	1.1 ± 0.5	99.9 ± 0.1	50.4 ± 11.1
	experiment-2	77.8 ± 13.6	0.0 ± 0.0	12.6 ± 8.3	0.0 ± 0.0
	experiment-3	98.8 ± 0.4	0.0 ± 0.0	43.3 ± 10.3	0.0 ± 0.0
	experiment-4	100.0 ± 0.0	0.0 ± 0.0	95.0 ± 1.4	0.0 ± 0.0
	experiment-5	98.3 ± 0.5	0.0 ± 0.0	99.3 ± 0.4	0.0 ± 0.0