Table 1: Performance of models on COGITAO experiments for the *CompGen* and *EnvGen* studies across different experiment settings (ES). We report the ID (in-domain) and OOD (out-of-domain) results for Grid accuracy (i.e., percentage of perfect match) and the relative drop (Δ = ID - OOD) averaged over all the experiments within the experiment setting.

		Vanilla-ViT			**Grid-ViT**			**LLaD**	
Study	**ES**	**ID**	**OOD**	**Δ**	**ID**	**OOD**	**Δ**	**ID**	**OOD**
CompGen	C1	28.4	15.7	12.7	**71.1**	18.7	52.4	54.2	**26.2**
	C2	34.0	15.6	18.4	**69.0**	**18.6**	50.4	49.2	15.2
	C3	15.1	1.0	14.1	75.6	5.1	70.5	**78.9**	**6.4**
EnvGen	G1	98.3	76.7	21.6	99.2	**90.0**	9.2	**99.4**	89.1
	G2	68.1	0.0	68.1	97.0	**73.0**	24.0	**99.3**	64.3
	G3	57.1	17.5	39.6	**78.5**	**27.2**	51.3	76.9	**27.2**
	G4	47.6	21.3	26.3	75.2	13.8	61.4	**76.3**	**35.6**
	G5	70.0	0.0	70.0	**98.0**	0.2	97.8	71.0	**14.0**