Table 2: Evaluation of reasoning accuracy

80.2

83.7

- w/o perceive pretrain

- w/o vision planning

between methods on two metrics.			methods on three task domains.			
Method	Token ↑	Semantic ↑	Methodology	Blocks	Letters	
PAR	58.2	0.63	SuSIE (+oracle)	18.9	18.1	
EmbodiedGPT	65.9	0.68	Susin (+oracle)	10.9	10.1	

			_			
Method	Token ↑	Semantic ↑		Methodology	Blocks	
PAR	58.2	0.63	-	SuSIE (+oracle)	18.9	
<b>EmbodiedGPT</b>	65.9	0.68		CoTDiffusion	13.1	
PERIA (ours)	97.6	0.98		DEDIA (2000)	10.1	

0.83

0.79

Method	Token $\uparrow$	Semantic ↑	Methodology	Blocks	Letters	VIMA
PAR EmbodiedGPT	58.2 65.9	0.63 0.68	SuSIE (+oracle)	18.9	18.1	19.4
DEDIA	05.7	0.00	CoTDiffusion	13.1	15.8	17.6

PERIA (ours)

- w/o alignment

memous on three task domains.						
Methodology	Blocks	Letters	VIMA			
SuSIE (+oracle)	18.9	18.1	19.4			

Table 3: Comparisons of FID  $(\downarrow)$  between

**10.2** 

12.3

13.5

14.2

11.4

15.9