Table 1: Performance of STIC compared with the original LVLM model LLaVA-v.16 (Vicuna 13B) across vision-language reasoning tasks. Image data used for 13B model remain the same as what we used for the 7B model.

Model	LLaVA-Bench				MM-Vet							MMBench
	Complex	Conv	Detail	**All**	Rec	Ocr	Know	Gen	Spat	Math	**Total**	**All**
LLaVA-v1.6 (7B)	87.4	61.3	77.8	77.3	43.1	40.6	29.6	32.5	44.7	15.4	42.2	63.7
LLaVA-v1.6 (13B)	94.0	73.8	78.7	84.5	52.2	47.1	38.8	45.2	42.7	26.9	48.9	70.6
w/ STIC	93.5	**78.1**(+4.3)	**79.4**	**85.6**(+1.1)	**54.5**	**48.0**	**42.3**(+3.5)	**49.4**(+4.2)	42.0	23.1	**50.5**(+1.6)	**72.3**(+1.7)

Table 2: Test performance of `llava-v1.6-mistral-7b` using various prompts with DaR. We evaluate prompt quality using DaR as a prompting method. DaR=`None` represents the original LVLM model's performance. Normal prompt refers to the simple prompt we used for DaR in our paper. GPT-4's well-curated prompt refers to the prompt we used for preferred response generation, and we include Mistral 7B's curated prompt for additional comparison.

Model	DaR	LLaVA-Bench	MM-Vet	MMBench
LLaVA-v1.6 (7B)	None	77.3	42.2	63.7
	Normal Prompt	78.5(+1.2)	42.3(+0.1)	50.7(−13.0)
	Hallu Prompt	73.7(−3.6)	40.5(−1.7)	40.7(−23.0)
	Well-curated (Llama-3 8B)	77.2(+0.1)	40.0(−2.2)	60.1(−3.6)
	Well-curated (GPT-4)	79.1(+2.1)	42.9(+0.7)	60.9(−2.8)

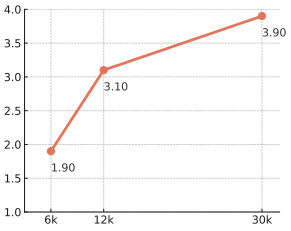

Figure 1: Scaling law for STIC (acc gain in LLaVA-Bench).

Table 3: Performance of STIC on different stage-1 training images compared with the original LVLM model LLaVA-v.16 (Vicuna 13B) across vision-language reasoning benchmarks.

Model	Data	LLaVA-Bench				MM-Vet							MMBench
		Complex	Conv	Detail	**All**	Rec	Ocr	Know	Gen	Spat	Math	**Total**	**All**
LLaVA-v1.6 (7B)	-	87.4	61.3	77.8	77.3	43.1	40.6	29.6	32.5	44.7	15.4	42.2	63.7
w/ STIC	COCO	89.1	63.7	79.5	79.2	45.7	42.5	30.4	34.9	45.1	22.7	45.0	67.8
w/ STIC	VFLAN	92.8	68.4	77.9	**81.9**	45.7	43.0	31.0	36.2	45.1	26.5	**45.1**	**68.3**

Query: I am getting one funghi pizza and one prosciutto pizza. How much should I pay in total?

Base (LLaVA-v1.6 7B):
Based on the menu prices, the total cost for one funghi pizza and one prosciutto pizza **would be 14,00€**.

STIC (LLaVA-v1.6 7B):
Based on the menu displayed in the image, **the price for a funghi pizza is 6,90€, and the price for a prosciutto pizza is 7,90€.** Therefore, the total cost for one funghi pizza and one prosciutto pizza **would be 14,80€**.

Figure 2: Response examples from original LLaVA-v1.6 and STIC (LLaVA-v1.6) in MM-Vet.

Query: What is the difference in metric tonnes between the amount of plastic produced and the amount discarded?

Base (LLaVA-v1.6 7B):
The difference between the amount of plastic produced and the amount discarded **is 302 million metric tonnes**.

STIC (LLaVA-v1.6 7B):
The image shows that in 2015, humans **produced 407 million metric tons of new plastic**, while **302 million metric tons were discarded as waste**. **The difference** between the amount of plastic produced and the amount discarded **is 105 million metric tons**. This indicates that a significant portion of the plastic produced is not being properly managed and is ending up in the environment or landfills.

Figure 3: Response examples from original LLaVA-v1.6 and STIC (LLaVA-v1.6) in MM-Vet.