LLMScore: Unveiling the Power of Large Language Models in Text-to-Image Synthesis Evaluation

Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, William Yang Wang

NEURAL INFORMATION PROCESSING SYSTEMS

UC SANTA BARBARA

UNIVERSITY of WASHINGTON

UNIVERSITY OF CALIFORNIA SANTA CRUZ

Motivation

Existing Metrics
- Fail to distinguish object-level alignment
- Single-aspect
- Non-interpretable

Text Prompt: A red book and a yellow vase.

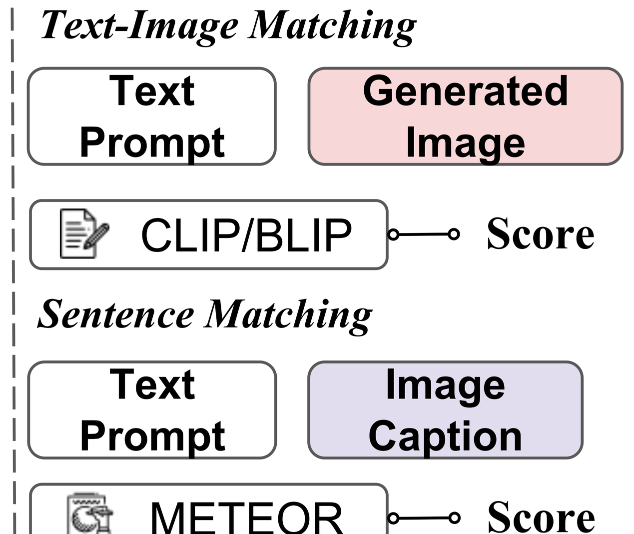
		Left	Right
Human	Overall	1.00	0.45
	Error Counting	1.00	0.55
Baseline	CLIP	0.27	0.31
	NegCLIP	0.26	0.32
	BLIP-ITM	0.99	1.00
	BLIP-ITC	0.48	0.49
LLMScore	Overall	1.00	0.50
	Error Counting	1.00	0.44

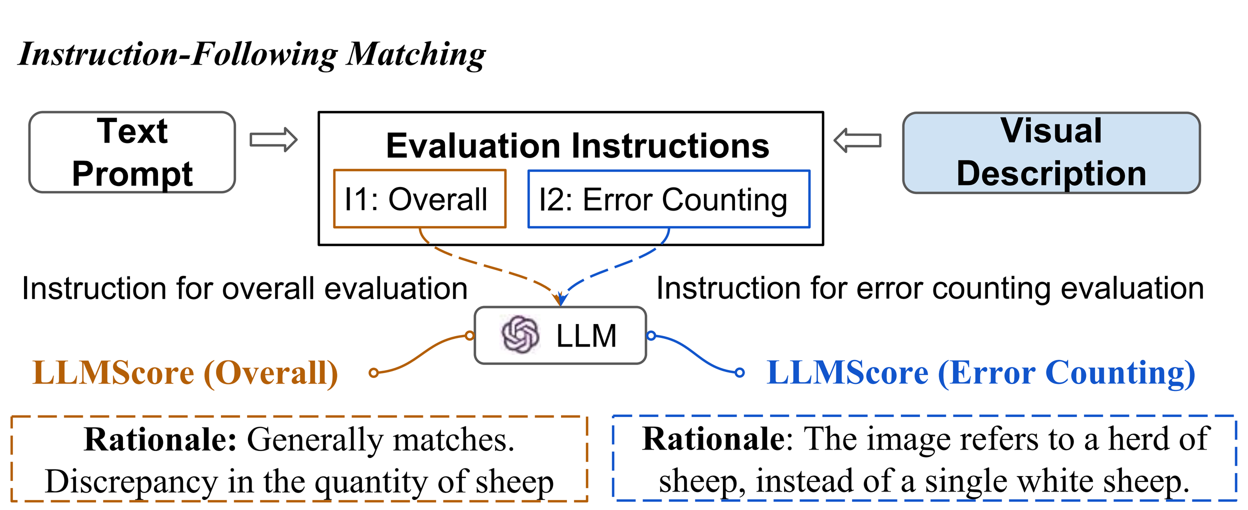
LLMScore - Overall Rationale
The overall quality of the image is quite low due to the significant discrepancies between the objects described in the text prompt and those portrayed in the image.

LLMScore - Error Counting Rationale
The red book from the prompt is not in the image. The vase is described as red in the image, while the text prompt specified a yellow vase. Over-specification of the yellow flowers in the image.

Paradigm Comparison

A red car and a white sheep.

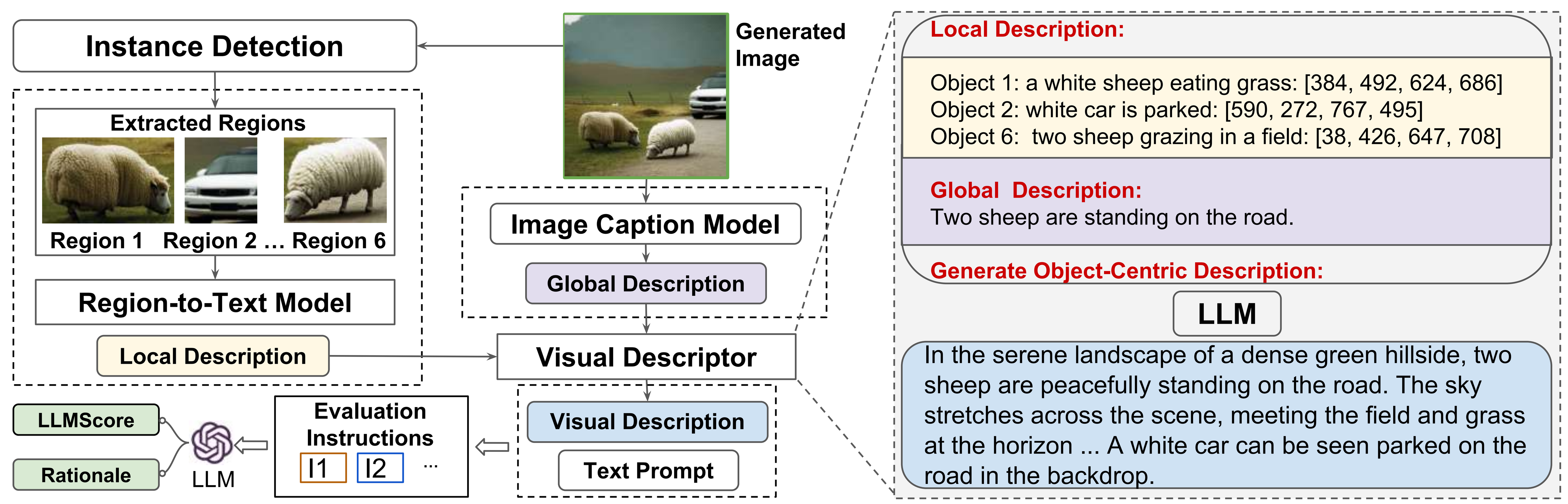




LLMScore Pipeline for Text-to-Image Evaluation

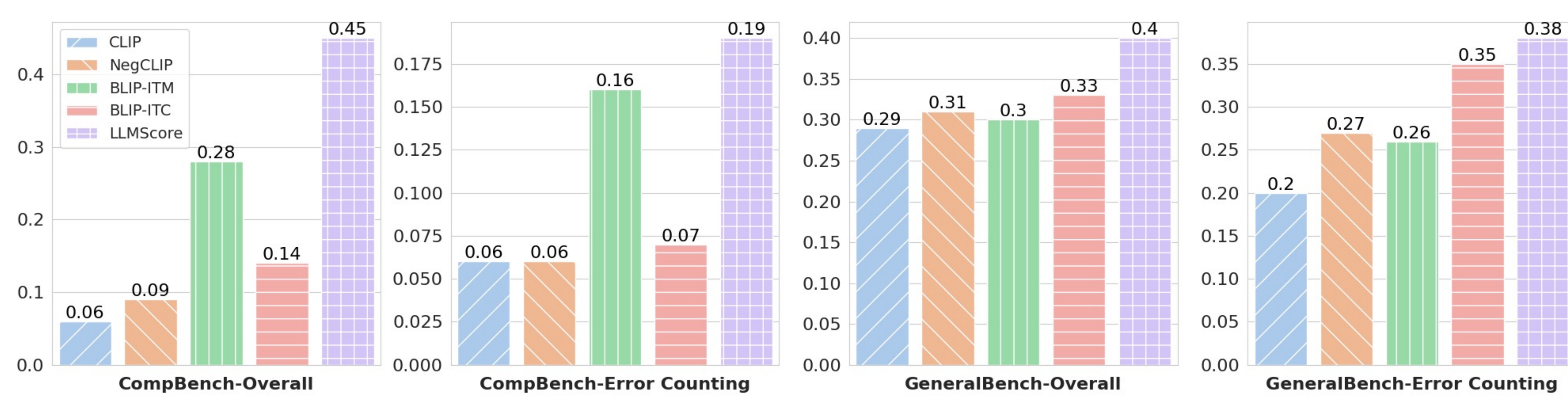
LLMs As Multi-Granularity Visual Descriptor
- Global Image Description
- Local Region Descriptions
- Object-Centric Visual Descriptions

LLMs As Text-to-Image Evaluator
- Instruction Following Rating
- Generating Rationale



Main Results

Averaged Kendall's τ Ranking Correlation with Human Ratings.



Example showing the LLMScore captures the object-level discrepancies.

A red clock and a gold suitcase.

Human	Overall	0.33
	Error Counting	0.70
Baseline	CLIP	0.30
	NegCLIP	0.36
	BLIP-ITM	0.99
	BLIP-ITC	0.45

LLMScore (Overall) 0.40

Overall Rationale
The image caption describes two red suitcases, a vintage clock on the wall, and a blue curtain in the background, but the text prompt only mentions a red clock and a gold suitcase. The alignment between the text prompt and image caption is weak.

LLMScore (Error Counting) 0.55

Error Counting Rationale
The composition errors include incorrect suitcase colors, incorrect clock color, and additional elements not mentioned in the text prompt (a second suitcase and blue curtain).

Ablation

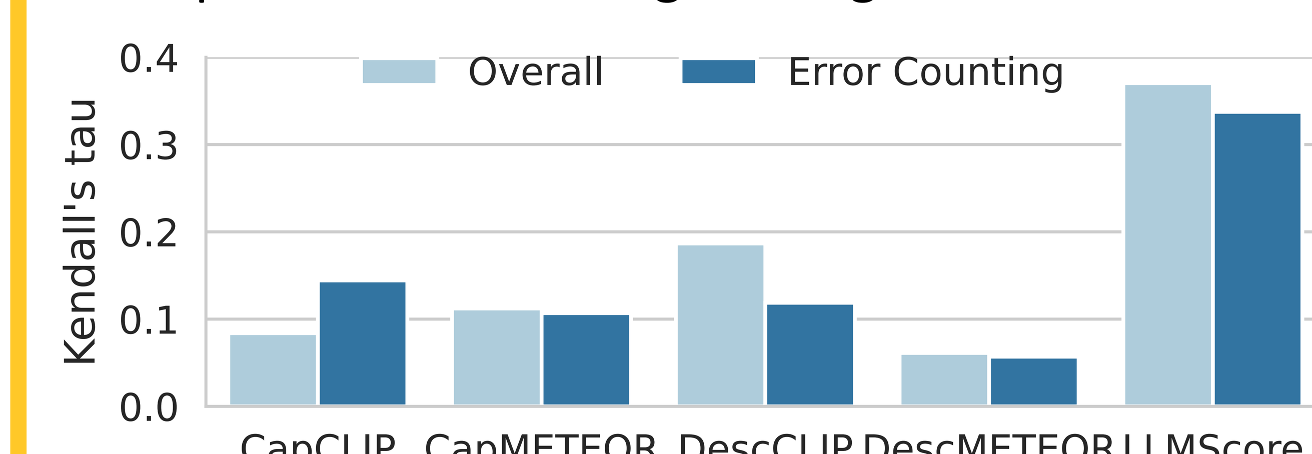
Composition-focused Prompt Bench

Human	Metric	Concept Conjunction				Attribute Binding Contrast			
		Stable Diffusion		DALLE		Stable Diffusion		DALLE	
		$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$
Overall	CLIP	0.1698	0.2459	−0.0049	−0.0058	0.0186	0.0320	0.0396	0.0548
	NegCLIP	0.1724	0.2504	0.0682	0.0995	0.0151	0.0211	0.1145	0.1634
	BLIP-ITM	0.4058	0.5618	0.3768	0.5266	0.1799	0.2559	0.1500	0.2134
	BLIP-ITC	0.2378	0.3398	0.0991	0.1413	0.1982	0.2814	0.0252	0.0344
	LLMScore	**0.4871**	**0.6956**	**0.5167**	**0.7230**	**0.4005**	**0.5480**	**0.3955**	**0.5506**
Error Counting	CLIP	0.2012	0.2864	−0.0782	−0.1107	0.0061	0.0071	0.0914	0.1286
	NegCLIP	0.2245	0.3240	−0.0353	−0.0502	−0.0339	−0.0418	0.0796	0.1130
	BLIP-ITM	0.3341	0.4561	0.1105	0.1668	0.0696	0.0968	0.1249	0.1783
	BLIP-ITC	0.2210	0.3124	−0.0755	−0.1071	0.0895	0.1315	0.0533	0.0786
	LLMScore	**0.3779**	**0.5443**	**0.2880**	**0.4428**	**0.1863**	**0.2821**	**0.2326**	**0.3351**

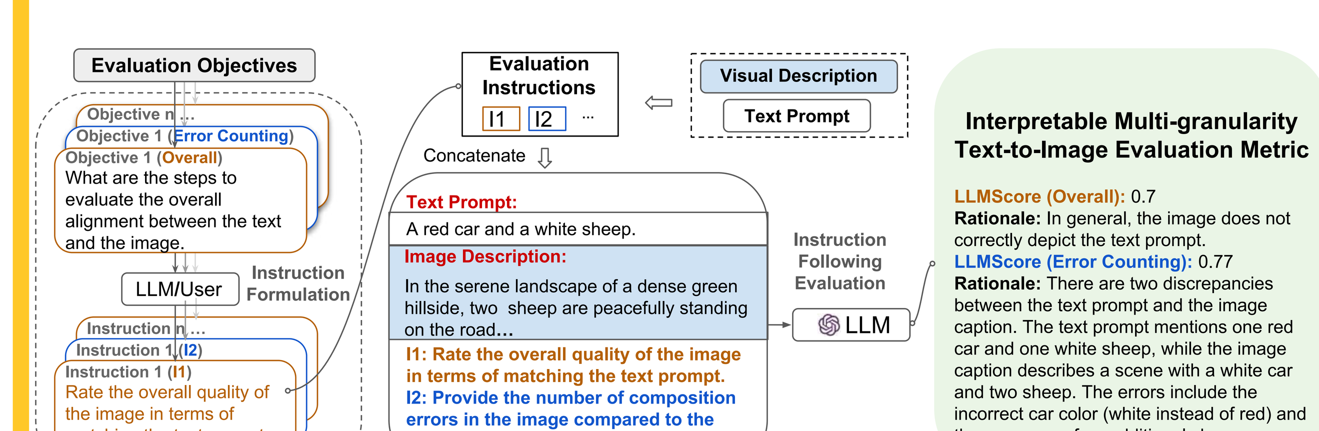
General-purpose Prompt Bench

Human	Metric	COCO2014		COCO2017		DrawBench		PaintSkills	
		$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$
Overall	CLIP	0.1971	0.2655	0.2227	0.2771	0.1530	0.2143	0.4715	0.5869
	NegCLIP	0.2164	0.2905	0.2793	0.3523	0.1463	0.1999	0.4911	0.6313
	BLIP-ITM	0.3252	0.4255	0.0928	0.1155	0.1044	0.1455	0.4755	0.6214
	BLIP-ITC	0.3465	0.4535	0.1703	0.2121	0.1569	0.2171	0.4743	0.5864
	LLMScore	**0.3629**	**0.4612**	**0.3357**	**0.4275**	**0.2230**	**0.3023**	**0.5600**	**0.6853**
Error Counting	CLIP	0.1464	0.2142	0.1888	0.2677	0.1360	0.1910	0.3052	0.2891
	NegCLIP	0.2116	0.3061	0.1795	0.2581	0.1179	0.1596	0.4563	0.4908
	BLIP-ITM	0.2251	0.3289	0.1137	0.1635	0.0871	0.1189	0.4622	0.4997
	BLIP-ITC	0.2636	0.3739	0.1849	0.2620	0.1506	0.2029	0.6178	0.6511
	LLMScore	**0.2830**	**0.3992**	**0.2038**	**0.3027**	**0.2134**	**0.2865**	**0.6437**	**0.7325**

Comparison of Matching Paradigms



Effects of Large Language Models



Human	LLM	DescCLIP		DescMeteor		LLMScore	
		$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$
Overall	GPT-3.5	0.1479	0.1956	0.0042	0.0073	0.2480	0.3285
	GPT-4	0.1128	0.1485	0.0297	0.0374	0.2793	0.3649
Error Counting	GPT-3.5	0.0467	0.0670	−0.0597	−0.0835	0.2205	0.3013
	GPT-4	0.0149	0.0228	−0.1087	−0.1494	0.2131	0.2981