MMIG-Bench: Towards Comprehensive and Explainable Evaluation of Multi-Modal Image Generation Models

*Hang Hua *, Ziyun Zeng *, Yizhi Song *, Yunlong Tang, Liu He, Daniel Aliaga, Wei Xiong, Jiebo Luo*

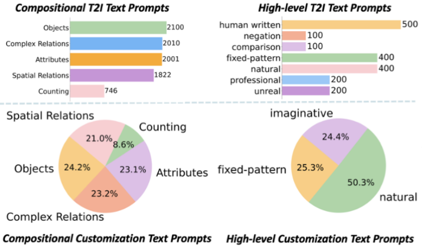
*Equal Contribution

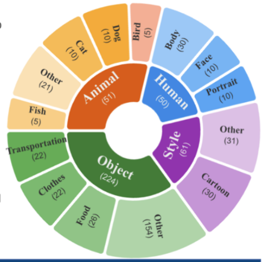
UNIVERSITY of ROCHESTER · NVIDIA · NEURAL INFORMATION PROCESSING SYSTEMS

Contribution

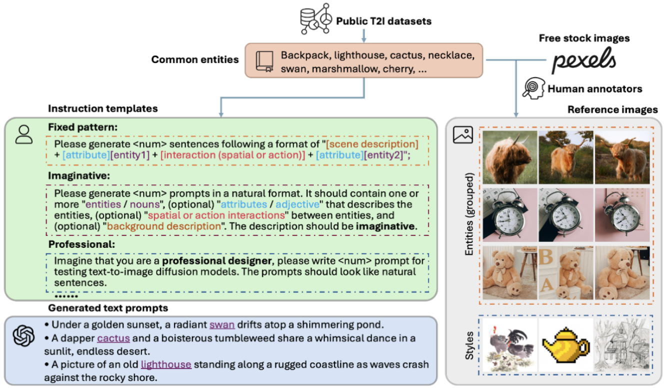
Unified task coverage and multi-modal input. We collect over 380 groups (animal, object, human, and style) comprising 1,750 multiview object-centric images enabling rigorous reference-based generation. We also construct 4,850 richly annotated prompts across compositionality (attributes, relations, objects, and numeracy), style (fixed pattern, professional, natural, human-written), realism (imaginative) and common sense (comparisons, negations).

Three-level evaluation suite. We propose a multilevel scoring framework for comprehensive evaluation. (1) Low-level metrics assess visual artifacts and identity preservation of objects; (2) At mid-level, we propose the **Aspect Matching Score (AMS)** : a novel VQA-based metric that captures fine-grained semantic alignment, showing strong correlation with human perception; (3) high-level metrics measure aesthetics and human preferences.





Data Curation



Backpack, lighthouse, cactus, necklace, swan, marshmallow, cherry, ...

Fixed pattern:

Please generate <num> sentences following a format of "[scene description] + [attribute][entity1] + [interaction (spatial or action)] + [attribute][entity2]";

Imaginative:

Please generate <num> prompts in a natural format. It should contain one or more "entities / nouns", (optional) "attributes / adjective" that describes the entities, (optional) "spatial or action interactions" between entities, and (optional) "background description". The description should be **imaginative**.

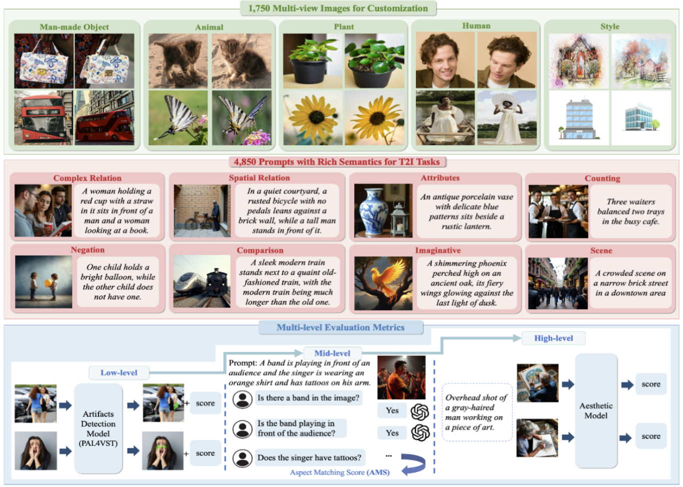
Professional:

Imagine that you are a **professional designer**, please write <num> prompt for testing text-to-image diffusion models. The prompts should look like natural sentences.

......

- Under a golden sunset, a radiant swan drifts atop a shimmering pond.
- A dapper cactus and a boisterous tumbleweed share a whimsical dance in a sunlit, endless desert.
- A picture of an old lighthouse standing along a rugged coastline as waves crash against the rocky shore.

Overview of MMIG-Bench


1,750 Multi-view Images for Customization
Man-made Object
Animal
Plant
Human
Style
4,850 Prompts with Rich Semantics for T2I Tasks
Complex Relation
Spatial Relation
Attributes
Counting
Negation
Comparison
Imaginative
Scene
Multi-level Evaluation Metrics
Low-level
Mid-level
High-level
Artifacts Detection Model (PAL4VST)
score
Yes
Yes
...
Aspect Matching Score (AMS)
Aesthetic Model

Complex Relation: A woman holding a red cup with a straw in it sits in front of a man and a woman looking at a book.

Spatial Relation: In a quiet courtyard, a rusted bicycle with no pedals leans against a brick wall, while a tall man stands in front of it.

Attributes: An antique porcelain vase with delicate blue patterns sits beside a rustic lantern.

Counting: Three waiters balanced two trays in the busy cafe.

Negation: One child holds a bright balloon, while the other child does not have one.

Comparison: A sleek modern train stands next to a quaint old-fashioned train, with the modern train being much longer than the old one.

Imaginative: A shimmering phoenix perched high on an ancient oak, its fiery wings glowing against the last light of dusk.

Scene: A crowded scene on a narrow brick street in a downtown area

Prompt: *A band is playing in front of an audience and the singer is wearing an orange shirt and has tattoos on his arm.*

Is there a band in the image?

Is the band playing in front of the audience?

Does the singer have tattoos?

Overhead shot of a gray-haired man working on a piece of art.

Performance on Text-to-Image Generation Tasks

Method	Low Level		Mid Level		High Level		
	CLIP-T ↑	PAL4VST ↓	AMS ↑	Human ↑	Aesthetic ↑	HPSv2 ↑	PickScore ↑
Diffusion Models							
SDXL [41]	33.529	14.340	79.08	72.29	6.337	0.277	0.120
Photon-v1 [40]	33.296	2.947	77.12	69.49	6.391	0.284	0.088
Lumina-2 [42]	33.281	15.531	84.11	73.18	6.048	0.287	0.116
HunyuanDit-v1.2 [31]	33.701	8.024	83.61	74.89	6.379	0.300	0.144
Pixart-Sigma-xl2 [2]	33.682	9.283	83.18	76.65	6.409	0.304	0.165
Flux.1-dev [25]	33.017	2.171	84.44	76.44	6.433	0.307	0.210
SD 3.5-large [6]	33.873	6.359	85.33	77.04	6.318	0.294	0.157
HiDream-I1-Full [50]	**33.876**	**1.522**	**89.65**	**83.18**	**6.457**	**0.321**	**0.450**
Autoregressive Models							
JanusFlow [33]	31.498	365.663	70.25	75.69	5.221	0.209	0.031
Janus-Pro-7B [3]	33.358	31.954	85.35	80.36	6.038	0.275	0.129
API-based Models							
Gemini-2.0-Flash [11]	32.433	11.053	85.35	81.98	6.102	0.275	0.110
GPT-4o [35]	32.380	3.497	82.57	81.02	6.719	0.279	0.263

Visit our GitHub repo for code and data

https://github.com/hanghuacs/MMIG-Bench

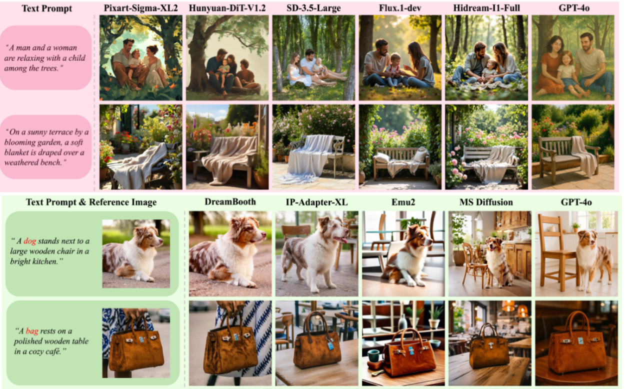
Performance on Customization Tasks

Method	Low Level					Mid Level		High Level		
	CLIP-T ↑	CLIP-I ↑	DINOv2 ↑	CUTE ↑	PAL4VST ↓	BLIPVQA ↑	AMS ↑	Aesthetic ↑	HPSv2 ↑	PickScore ↑
Diffusion Models										
BLIP Diffusion[29]	26.137	80.286	26.232	69.681	56.780	0.247	41.59	5.830	0.213	0.032
DreamBooth [45]	24.227	**88.758**	**38.961**	**79.780**	43.535	0.108	28.00	5.368	0.179	0.019
Emu2 [48]	28.410	79.026	31.831	71.132	10.461	0.378	53.13	5.639	0.243	0.066
Ip-Adapter-XL [60]	28.577	85.297	34.177	74.995	8.531	0.290	51.10	5.840	0.233	0.073
MS Diffusion [54]	31.446	77.827	23.600	71.306	4.748	0.496	71.40	5.979	0.271	0.143
API-based Models										
GPT-4o [35]	**33.527**	75.152	25.174	64.776	**1.973**	**0.672**	**90.90**	**6.368**	**0.289**	**0.550**

Aspect Matching Score (AMS)

Method	BLIPVQA ↑	VQ2 ↑	DSG ↑	AMS ↑	Human ↑
Diffusion Models					
SDXL	0.433	69.07	87.63	79.08	72.29
Photon-v1	0.440	66.84	86.26	77.12	69.49
Lumina-2	0.517	72.51	90.12	84.11	73.18
HunyuanDiT-v1.2	0.513	73.13	89.77	83.61	74.89
Pixart-Sigma-xl2	0.521	71.51	89.69	83.18	76.65
Flux.1-dev	0.511	71.41	83.33	84.44	76.44
SD 3.5-large	0.525	73.28	91.41	85.33	77.04
HiDream-I1-Full	0.572	75.09	92.43	89.65	83.18
Autoregressive Models					
JanusFlow [33]	0.390	57.24	85.43	70.25	75.69
Janus-Pro [3]	0.530	67.41	92.15	85.35	80.36
API-based Models					
Gemini-2.0-Flash	0.495	72.01	92.93	85.40	81.98
GPT-4o	0.497	70.34	89.99	82.57	81.02

Case Study



"A man and a woman are relaxing with a child among the trees."

"On a sunny terrace by a blooming garden, a soft blanket is draped over a weathered bench."

" A dog stands next to a large wooden chair in a bright kitchen."

"A bag rests on a polished wooden table in a cozy café."