AesGI-Bench: Benchmarking and Evaluating the Aesthetic Quality of AI-Generated Images via Large Multimodal Models

Jiarui Wang*, Xubo Su*, Huiyu Duan, Weize Sun, Ziheng Jia, Juntong Wang, Yu Zhao, Guangtao Zhai, Xiongkuo Min

Shanghai Jiao Tong University, Shanghai, China * Equal contribution

NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2026

1. Introduction

➢ Modern T2I models produce plausible images. **Fine-grained aesthetic evaluation** remains challenging.

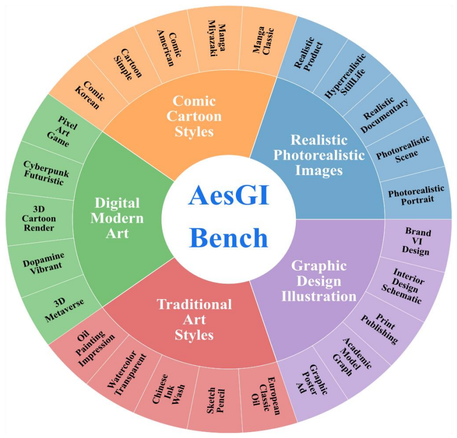
19,485 AI-generated images

117,672 pairwise comparisons

18 generation models

1,249 curated prompts

5 styles and 25 subcategories



Prompt construction & annotation

➢ Prompts from **COCO and DiffusionDB**, refined with **Qwen3.5**.

➢ **30 annotators** study the annotation guidelines before labeling. The **collected annotations** then undergo quality filtering.

2. Dataset: AesGI-Bench



Visual Aesthetic

Color harmony, composition, lighting and visual appeal

Technical Quality

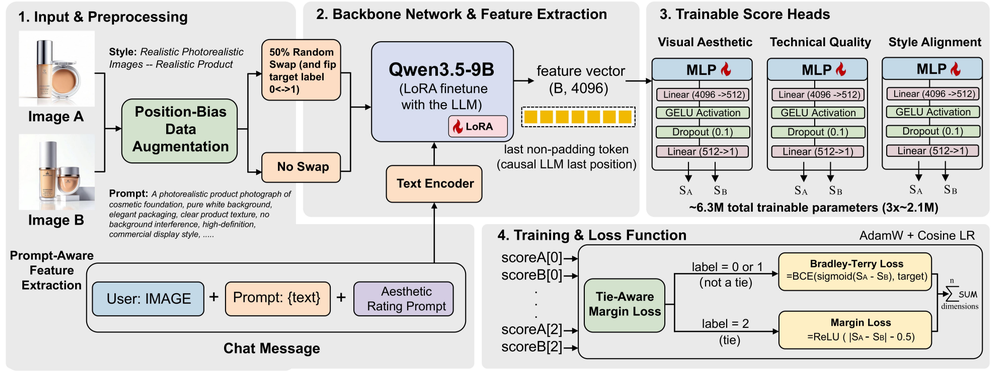
Texture, structural clarity and absence of artifacts

Style Alignment

Consistency with the style specified in the prompt

➢ Human annotators compare images from the same prompt across **three aesthetic dimensions**.

3. Model: AesGI-Assessor



➢ **Qwen3.5-9B** with a **frozen vision encoder** and **language-side LoRA** (~3.9M parameters).

➢ **Three MLP heads** (~6.3M parameters) predict one score per dimension for **each image**.

➢ **Bradley-Terry loss** handles win/loss pairs. **Tie-aware margin loss** handles ties (m = 0.5).

4. Main results

➢ Evaluation on AesGI-Bench

Table 2. Performance comparison of automatic quality evaluation models on the AesGI-Bench.

Method	Visual Aesthetic			Technical Quality			Style Alignment			Overall		
Dimensions / Metrics	SRCC↑	PLCC↑	KRCC↑	SRCC↑	PLCC↑	KRCC↑	SRCC↑	PLCC↑	KRCC↑	SRCC↑	PLCC↑	KRCC↑
♠ BMPRI [30]	0.238	0.517	0.150	0.318	0.555	0.170	0.150	0.458	0.052	0.230	0.513	0.137
♠ BPRI [29]	0.166	0.429	0.111	0.209	0.474	0.131	0.160	0.371	0.092	0.154	0.428	0.098
♠ BRISQUE [28]	0.358	0.485	0.307	0.393	0.531	0.315	0.239	0.407	0.210	0.356	0.478	0.294
♠ HOSA [48]	0.209	0.389	0.124	0.299	0.451	0.210	0.106	0.302	0.052	0.205	0.384	0.111
♠ NIQE [27]	0.313	0.440	0.229	0.360	0.496	0.262	0.182	0.373	0.131	0.311	0.440	0.216
♠ QAC [49]	0.325	0.463	0.255	0.339	0.498	0.236	0.241	0.399	0.184	0.329	0.456	0.268
♡ CLIPScore [31]	0.300	0.382	0.163	0.301	0.389	0.170	0.405	0.470	0.236	0.325	0.415	0.176
♡ BLIPScore [32]	0.139	0.197	0.059	0.092	0.186	0.013	0.263	0.308	0.131	0.162	0.230	0.072
♡ Imagereward [33]	0.496	0.576	0.359	0.507	0.588	0.341	0.516	0.618	0.354	0.511	0.597	0.373
♡ HPSv2.1 [20]	0.501	0.601	0.333	0.554	0.642	0.367	0.436	0.591	0.289	0.494	0.614	0.320
♡ PickScore [21]	0.725	0.738	0.569	0.749	0.772	0.603	0.659	0.730	0.472	0.721	0.750	0.556
♣ LIQE [50]	0.174	0.290	0.137	0.277	0.348	0.184	0.080	0.224	0.039	0.160	0.290	0.124
♣ MUSIQ [51]	0.199	0.287	0.124	0.291	0.348	0.197	0.138	0.226	0.066	0.195	0.290	0.111
♣ AestheticScore [16]	0.370	0.510	0.229	0.436	0.567	0.289	0.243	0.419	0.131	0.354	0.502	0.216
♣ ROC4MLLM [14]	0.387	0.526	0.216	0.456	0.591	0.302	0.275	0.450	0.118	0.375	0.526	0.203
♣ QAlign [34]	0.558	0.704	0.346	0.626	0.758	0.433	0.468	0.654	0.275	0.552	0.709	0.333
♣ Artimuse [15]	0.763	0.828	0.595	0.758	0.856	0.577	0.657	0.765	0.498	0.746	0.821	0.582
♣ MINTIQA-Quality [35]	0.628	0.747	0.425	0.687	0.783	0.459	0.545	0.711	0.380	0.622	0.751	0.412
♣ MINTIQA-Authenticity [35]	0.695	0.803	0.503	0.735	0.839	0.511	0.608	0.762	0.459	0.695	0.806	0.490
♣ MINTIQA-Correspondence [35]	0.775	0.829	0.582	0.799	0.866	0.590	0.664	0.782	0.485	0.773	0.830	0.569
△ mPLUG-Owl3 (7B) [52]	0.480	0.502	0.268	0.690	0.783	0.498	0.464	0.674	0.289	0.515	0.718	0.307
△ Llava-one-vision (8B) [53]	0.507	0.652	0.359	0.665	0.789	0.459	0.469	0.569	0.315	0.548	0.733	0.386
△ GLM-4.1V (9B) [54]	0.577	0.630	0.438	0.622	0.712	0.485	0.706	0.779	0.577	0.695	0.768	0.556
△ InternVL3 (8B) [55]	0.424	0.523	0.294	0.376	0.439	0.275	0.722	0.772	0.564	0.470	0.648	0.268
△ InternVL3.5 (8B) [56]	0.695	0.738	0.503	0.636	0.710	0.472	0.736	0.697	0.551	0.711	0.783	0.503
△ InternVL3.5 (38B) [56]	0.684	0.720	0.491	0.689	0.776	0.507	0.736	0.697	0.551	0.756	0.837	0.595
△ VILA (8B) [57]	0.610	0.782	0.451	0.724	0.782	0.538	**0.812**	**0.847**	**0.630**	0.664	0.819	0.516
△ Qwen2.5-VL (7B) [58]	0.657	0.751	0.464	0.722	0.787	0.525	0.788	0.764	0.616	0.736	0.819	0.556
△ Qwen3.5 (9B) [38]	0.820	0.757	0.621	**0.882**	**0.828**	**0.734**	0.749	0.721	0.577	**0.870**	**0.852**	0.673
△ Qwen3.5 (35B) [38]	**0.851**	**0.835**	**0.660**	0.733	0.811	0.538	0.721	0.694	0.551	0.868	0.847	**0.699**
△ **AesGI-Assessor (Ours)**	**0.866**	**0.863**	**0.752**	**0.963**	**0.834**	**0.866**	**0.845**	**0.850**	**0.695**	**0.891**	**0.853**	**0.778**

➢ Zero-shot Cross-dataset Evaluation

Table 3. Zero-shot Performance comparisons on the EvalMi-50K from perception dimension.

Dimension	Image Level			Model Level		
Methods / Metrics	SRCC	PLCC	KRCC	SRCC	PLCC	KRCC
♠ NIQE [27]	0.3818	0.3885	0.2589	0.6922	0.7974	0.5217
♠ QAC [49]	0.0376	0.0855	0.0246	0.0035	0.1209	0.0072
♠ BRISQUE [28]	0.0157	0.0334	0.0104	0.0269	0.0289	0.0003
♠ BPRI [29]	0.0329	0.0196	0.0207	0.0052	0.0007	0.0002
♠ HOSA [48]	0.1480	0.1690	0.0985	0.4626	0.4694	0.3405
♠ BMPRI [30]	0.1519	0.1245	0.1011	0.3913	0.3549	0.2754
♠ Higrade-2 [59]	0.0393	0.0260	0.0275	0.1896	0.1086	0.0869
♣ MUSIQ [51]	0.4519	0.5215	0.3070	0.6835	0.8183	0.5289
♣ AestheticScore [16]	0.3771	0.4433	0.2534	0.6583	0.6204	0.4855
♣ Artimuse [15]	0.3639	0.3932	0.2474	0.4833	0.7533	0.3333
♣ MINTIQA-Quality [35]	0.4078	0.4257	0.2912	0.8265	**0.8251**	0.6522
♣ MINTIQA-Authenticity [35]	0.3954	0.4099	0.2714	0.8226	0.8186	0.6539
♣ MINTIQA-Correspondence [35]	0.4096	0.4342	0.2813	0.7939	0.8122	0.6232
△ LLaVA-1.5 (7B) [60]	0.3372	0.3525	0.2577	0.5215	0.5301	0.3333
△ LLaVA-NeXT (8B) [61]	0.4333	0.4164	0.3442	0.7113	0.7224	0.5449
△ mPLUG-Owl3 (7B) [52]	0.3918	0.3569	0.3018	0.7052	0.7343	0.5239
△ MiniCPM-V2.6 (8B) [62]	0.3733	0.1053	0.2839	0.6548	0.7021	0.4855
△ Qwen2-VL (7B) [63]	0.3760	0.3625	0.3061	0.7826	0.7830	0.6029
△ DeepSeekVL (7B) [64]	0.2611	0.3010	0.1988	0.2356	0.3457	0.1872
△ CogAgent (18B) [65]	0.3861	0.4235	0.2927	0.7739	0.7727	0.6094
△ InternVL2.5 (8B) [66]	0.2597	0.3669	0.1859	0.7635	0.7356	0.6014
△ IntLM-XC (7B) [67]	0.3918	0.3569	0.3018	0.6122	0.6746	0.4493
△ Qwen3.5 (9B) [38]	**0.5333**	**0.5414**	**0.4058**	**0.8304**	0.7973	**0.6754**
△ Qwen3.5 (35B) [38]	0.4888	0.4657	0.3873	0.7826	0.7595	0.6246
△ **AesGI-Assessor (Ours)**	**0.6618**	**0.6308**	**0.4726**	**0.8461**	**0.8349**	**0.6957**

➢ Main findings

➢ AesGI-Assessor achieves the **highest overall correlation** with human judgments on AesGI-Bench.

➢ The model transfers to EvalMi-50K **without additional fine-tuning**, improving agreement at both **image and model levels**.