WORLD LAUREATES FORUM
世界顶尖科学家论坛
2026
国际工程智能大会
2026 International Conference on AI for Engineering
World Laureates Foundation
上海世界顶尖科学家发展基金会
工程智能研究院
INSTITUTE OF AI FOR ENGINEERING

T2I-CoReBench: Composition and Reasoning Evaluation for Text-to-Image Generation

Ouxiang Li[1*], Yuan Wang[1], Xinting Hu[1†], Huijuan Huang[2‡], Rui Chen[2], Jiarong Ou[2], Xin Tao[2†], Pengfei Wan[2], Xiaojuan Qi[3], Fuli Feng[1]
[1] University of Science and Technology of China [2] Kling Team, Kuaishou Technology
[3] The University of Hong Kong
† Corresponding authors ‡ Project lead * Work done during internship at Kling Team, Kuaishou Technology

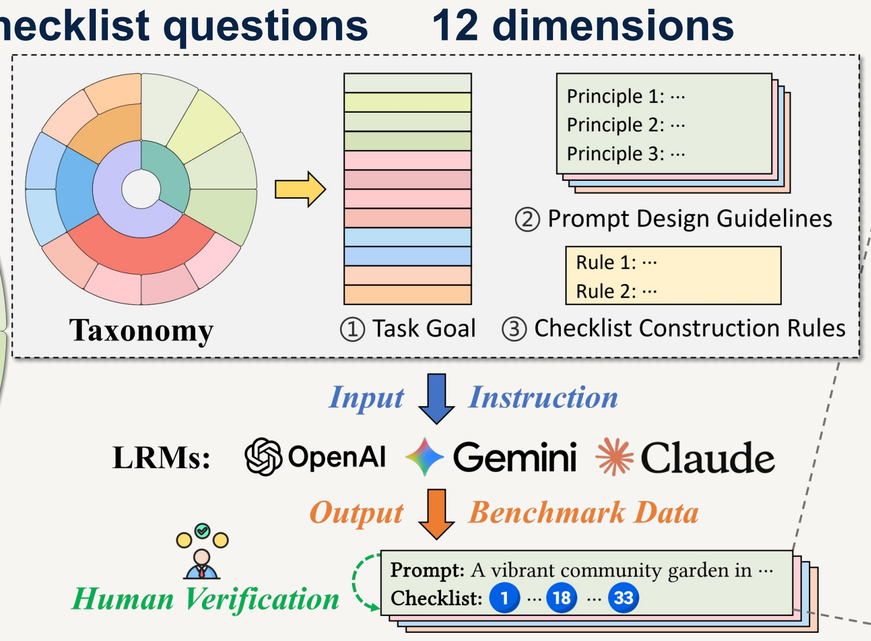
01 BENCHMARK AND EVALUATION

Explicit composition and implicit reasoning in text-to-image generation

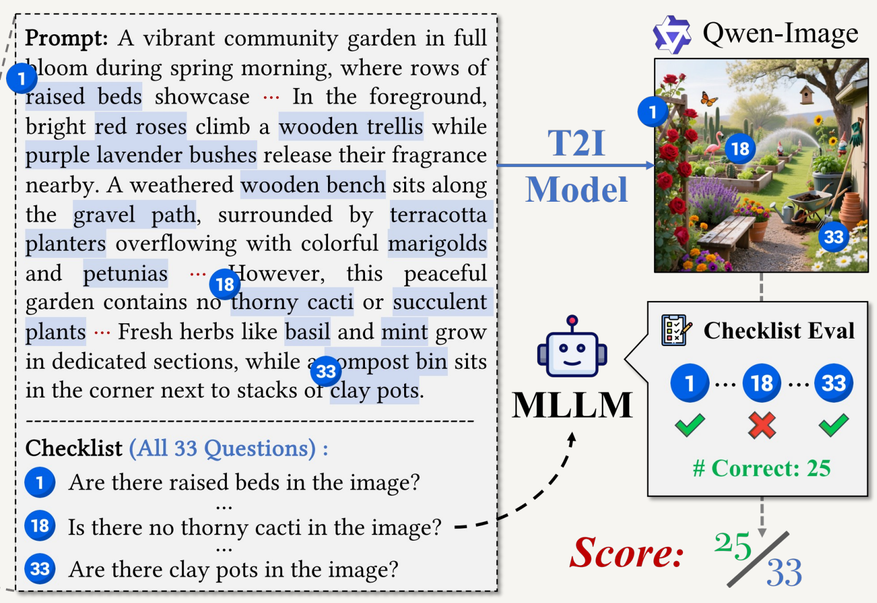
1,080 prompts **13,536 checklist questions** **12 dimensions**



4 composition + 8 reasoning dimensions



(a) Data Generation



(b) Evaluation Protocol

Score = satisfied checklist items / all items. Prompts and checklists are human-verified.

02 EVALUATION RESULTS: ALL 38 MODELS

Gemini 2.5 Flash. Mean scores (%) from Table 3. Bold: best overall in each group. Gold rows: closed-source models

Model	Composition	Reasoning	Overall
SD-3-Medium	40.4	34.5	36.5
SD-3.5-Medium	41.0	33.9	36.3
SD-3.5-Large	41.5	35.9	37.8
FLUX.1-schnell	49.6	40.0	43.2
FLUX.1-dev	48.6	39.0	42.2
FLUX.1-Krea-dev	56.0	44.3	48.2
FLUX.2-dev	**84.7**	**54.2**	**64.4**
FLUX.2-klein-4B	68.4	47.7	54.6
FLUX.2-klein-9B	78.0	52.0	60.6
PixArt-α	25.0	23.7	24.1
PixArt-Σ	30.9	28.5	29.3
HiDream-I1	50.3	39.4	43.0
Qwen-Image	78.0	49.3	58.9
Qwen-Image-2512	83.7	51.7	62.4
Z-Image-Turbo	74.6	47.3	56.4
LongCat-Image	70.8	54.1	59.6
Infinity-8B	46.4	42.3	43.6
GoT-R1-7B	35.8	31.2	32.7
BAGEL	46.4	34.1	38.2

Model	Composition	Reasoning	Overall
BAGEL w/ Think	39.6	41.9	41.1
show-o2-1.5B	40.1	33.2	35.5
show-o2-7B	40.4	34.6	36.5
Janus-Pro-1B	35.5	13.0	20.5
Janus-Pro-7B	40.5	19.8	26.7
BLIP3o-4B	28.5	28.2	28.3
BLIP3o-8B	30.3	31.8	31.3
OmniGen2-7B	49.9	39.4	42.9
HunyuanImage-3.0	78.9	48.6	58.7
Seedream 3.0	67.3	50.3	56.0
Seedream 4.0	86.1	69.9	75.3
Seedream 4.5	88.3	71.3	77.0
Gemini 2.0 Flash	62.1	51.2	54.8
Nano Banana	80.6	73.6	75.9
Nano Banana Pro	**89.7**	**82.7**	**85.0**
Imagen 4	78.4	65.9	70.0
Imagen 4 Ultra	82.4	72.9	76.1
GPT-Image	79.8	69.0	72.6
GPT-Image-1.5	85.5	74.6	78.2

03 EXAMPLE TASK AND COMPLETE CHECKLIST

Procedural reasoning: follow six origami steps and depict the final purple crane.



Qwen-Image: 4/5



Seedream 4.0: 5/5

Checklist criterion	Qwen-Image	Seedream 4.0
1 One final purple origami crane on the table	Yes	Yes
2 Both wings extended outward horizontally	Yes	Yes
3 Crane stands upright without external support	Yes	Yes
4 Distinct head bent slightly downward	Yes	Yes
5 No unfolded sheet or scrap paper on the table	No	Yes

PROJECT & PAPER

t2i-corebench.github.io
arxiv.org/abs/2509.03516

Paper GitHub Page

REFERENCE

Ouxiang Li et al. Easier Painting Than Thinking: Can Text-to-Image Models Set the Stage, but Not Direct the Play? ICLR 2026.