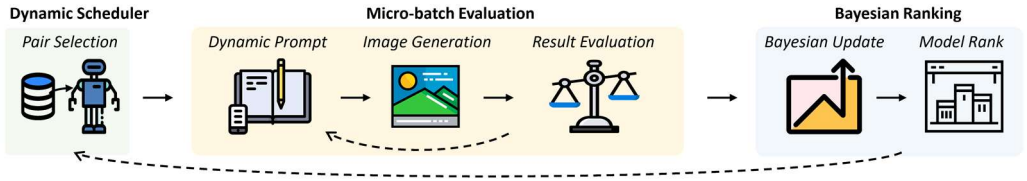
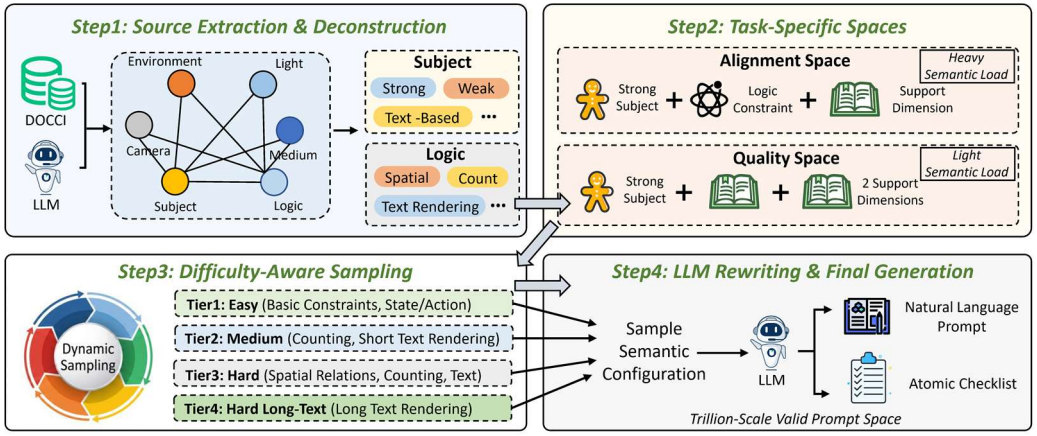
DynT2I-Eval: A Dynamic Evaluation Framework for Text-to-Image Models

Juntong Wang, Jiarui Wang, Huiyu Duan, Lewei Li, Guangtao Zhai, Xiongkuo Min
Shanghai Jiao Tong University, Shanghai, China



Overview of the dynamic prompt generation pipeline



Why Dynamic Evaluation?

Most text-to-image (T2I) benchmarks rely on fixed, publicly available prompt sets. Repeated exposure may encourage prompt-specific optimization, making benchmark scores less representative of open-world generalization.

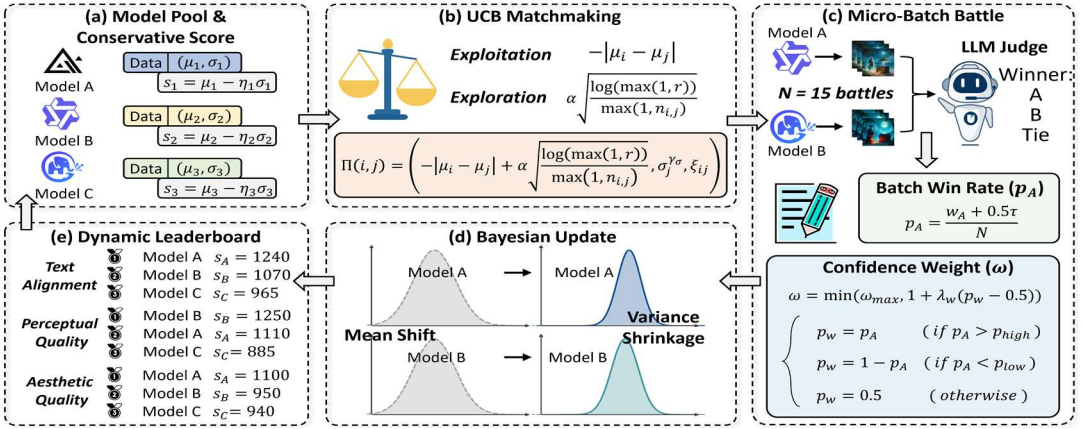
Our Goal: Build a continuously evolving evaluation framework that generates fresh, controllable prompts while maintaining reliable model rankings under changing test distributions.

Structured Prompt Space. We decompose 14,845 long-form image descriptions into six semantic dimensions, enabling approximately 10^{12} rule-valid configurations in each task-specific prompt space.

Difficulty-Aware Sampling. Four difficulty tiers provide controllable challenges, ranging from attribute binding and counting to spatial reasoning and long-text rendering.

Dynamic Ranking. Prompt-conditioned pairwise comparisons, adaptive match scheduling, micro-batch aggregation, and confidence-weighted Bayesian updates enable continuous leaderboard maintenance.

Overview of the dynamic ranking and scheduling framework



Model	Text Alignment				Perceptual Quality				Aesthetic Quality			
	Rk	s	μ	σ	Rk	s	μ	σ	Rk	s	μ	σ
FLUX.2-klein-9B [17]	1	1507.69	1617.64	36.65	4	1059.62	1142.07	27.49	9	565.02	653.12	29.37
LongCat-Image [34]	2	1361.83	1458.22	32.13	2	1407.11	1532.49	41.80	3	1198.11	1315.04	38.98
FLUX.2-klein-4B* [17]	3	1200.82	1368.95	56.04	8	883.49	1003.02	39.84	10	510.62	607.86	32.42
FLUX.1-Krea-dev [18]	4	1133.53	1239.13	35.20	9	733.53	821.74	29.40	1	1470.65	1579.26	36.20
Z-Image-Turbo [35]	5	1124.74	1232.08	35.78	1	1443.63	1576.85	44.41	12	347.66	446.19	32.84
SD-3.5-Large [6]	6	829.00	948.65	39.88	11	571.97	658.61	28.88	4	1183.99	1308.78	41.60
FLUX.1-dev [16]	7	771.57	868.78	32.40	3	1171.46	1288.40	38.98	5	949.07	1057.33	36.09
SD-3.5-Medium [6]	8	763.20	862.19	32.99	10	613.03	690.86	25.94	7	714.93	835.18	40.09
CogView4-6B [49]	9	567.48	662.80	31.77	12	567.45	651.46	28.00	11	462.32	555.52	31.07
SD-3-Medium [6]	10	550.83	637.25	28.81	7	905.71	1001.90	32.06	8	691.53	813.23	40.56
SD3.5-Large-Turbo [6]	11	483.74	599.04	38.43	6	994.74	1073.75	26.33	6	945.98	1051.20	35.07
SDXL-Base-1.0* [30]	12	-48.15	246.50	98.22	5	1019.97	1121.73	33.92	2	1405.49	1517.68	37.40

92.22% Relaxed agreement with human judgments across 450 pairwise evaluation cases.

80.66% Strict agreement with human judgments across the same 450 cases.

6 / 6 Simulated environments in which DynT2I-Eval achieves the best early-stage ranking quality at SRCC@500 and SRCC@1500.

100% Strong-model discovery success across all six simulated environments, requiring an average of 5.7–6.1 batches.