WorldGenBench: A World-Knowledge-Integrated Benchmark for Reasoning-Driven Text-to-Image Generation

Daoan Zhang[1*], Che Jiang[3*], Ruoshi Xu[3*], Biaoxiang Chen[3*], Zijian Jin[4], Yutian Lu[5]
Jianguo Zhang[3], Liang Yong[2], Jiebo Luo[1†], Shengda Luo[2,3†]

[1]University of Rochester, [2]Chinese Medicine Guangdong Laboratory, [3]Southern University of Science and Technology
[4] New York University, [5] Datawhale org.

daoan.zhang@rochester.edu, {12210914, xurs2022, 12112202 }@mail.sustech.edu.cn
zj2076@nyu.edu, physicoada@gmail.com, {zhangjg, luosd}@sustech.edu.cn
yongliangresearch@gmail.com, jluo@cs.rochester.edu

Abstract

Recent advances in text-to-image (T2I) generation have achieved impressive results, yet existing models still struggle with prompts that require rich world knowledge and implicit reasoning—both of which are critical for producing semantically accurate, coherent, and contextually appropriate images in real-world scenarios. To address this gap, we introduce ***WorldGenBench****, a benchmark designed to systematically evaluate T2I models' world knowledge grounding and implicit inferential capabilities, covering both the humanities and nature domains. We propose the* ***Knowledge Checklist Score****, a structured metric that measures how well generated images satisfy key semantic expectations. Experiments across 21 state-of-the-art models reveal that while diffusion models lead among open-source methods, proprietary auto-regressive models like GPT-4o exhibit significantly stronger reasoning and knowledge integration. Our findings highlight the need for deeper understanding and inference capabilities in next-generation T2I systems. Project Page: https://dwanzhang-ai.github.io/WorldGenBench/*

1. Introduction

Despite significant progress in text-to-image (T2I) generation [3, 11, 24, 25], most contemporary models excel primarily on prompts that involve explicit, surface-level descriptions. This reveals a fundamental limitation: their apparent success often results from direct pattern association rather than true understanding. However, generating high-quality, semantically accurate images in realistic and complex scenarios requires much more than simple lexical matching—it necessitates the ability to integrate **world knowledge** and perform **implicit reasoning**.

Figure 1. The presence of world knowledge and the emergence of implicit reasoning capabilities are fundamental to building a high-quality text-to-image model.

World knowledge is critical for interpreting references that are not exhaustively specified within the prompt but are assumed as common background information. For instance, understanding what a "medieval knight" should wear, what a "Victorian street" should look like, or the physical characteristics of a "polar landscape" relies on broad factual and commonsense grounding across domains such as history, culture, geography, and the physical sciences. Without access to such knowledge, models are prone to hallucinations, anachronisms, or incoherent scene compositions. Moreover, many prompts inherently require implicit reasoning—the capacity to infer unstated but logically necessary information based on minimal textual cues. For example, a prompt mentioning "a rainy soccer match" implicitly requires models to represent wet conditions, overcast skies,

*Equal Contribution
†Corresponding Author

and possibly slippery ground, even if these elements are not explicitly mentioned. In real-world scenarios, a model should be able to combine the prompt with relevant world knowledge, performing implicit reasoning to infer what elements must be present in the image for it to be coherent and contextually accurate. Failure to perform such reasoning leads to images that, while visually plausible in isolation, fail to semantically match the true intent of the prompt.

Therefore, to foster the development of T2I systems that can operate reliably in open-world settings, it is essential to move beyond superficial evaluations and systematically assess models' abilities in knowledge integration and inferential reasoning. Motivated by this, we present the first benchmark specifically designed to evaluate T2I models from the perspectives of world knowledge understanding and implicit reasoning capabilities. We further propose a more structured evaluation approach, we introduce the **Knowledge Checklist Score**. For each prompt, we construct a corresponding knowledge checklist to assess how many elements in the checklist are correctly reflected in the image. This approach significantly mitigates the hallucinations and inconsistencies caused by relying solely on subjective evaluations from VLMs, as seen in benchmarks like Wise [15].

2. Related Work

2.1. Evaluation of Text-to-Image Models

Traditional evaluation of text-to-image (T2I) models has focused on image realism and text-image alignment, using metrics like FID [9] and CLIPScore [8]. However, these approaches fall short in assessing a model's ability to understand and apply world knowledge. Recent benchmarks such as GenEval [7], T2I-CompBench [10], Commonsense-T2I [6], PhyBench [14], and Wise [15] introduce more challenging tasks involving compositionality, commonsense, and physical reasoning. Yet, they often rely on **simple world knowledge or explicit reasoning**, making these benchmarks quite different from real-world use cases.

3. WorldGenBench

A robust T2I model should demonstrate not only a comprehensive grasp of world knowledge but also strong implicit reasoning abilities. Specifically, it should be capable of generating user-expected and factually consistent content even when presented with incomplete or underspecified prompts, by effectively leveraging world knowledge and enabling the emergence of implicit inference. To rigorously assess these capacities, we introduce WorldGenBench, a benchmark specifically designed to evaluate T2I models' competencies in knowledge-grounded understanding and implicit reasoning. We illustrate the differences among various benchmarks in Table 1, comparing four T2I benchmarks—PhyBench, Commonsense-T2I, Wise, and WorldGenBench—across four key dimensions: World Knowledge, Implicit Reasoning, Culture-awareness, and Checklist Score. While most benchmarks address only one or two aspects, WorldGenBench stands out as the only one offering comprehensive coverage across all dimensions, making it a more robust and versatile tool for evaluating T2I systems.

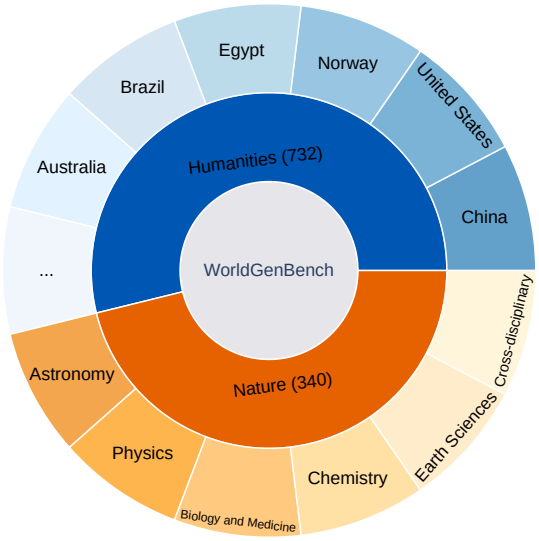


Figure 2. Detailed static information of WorldGenBench.

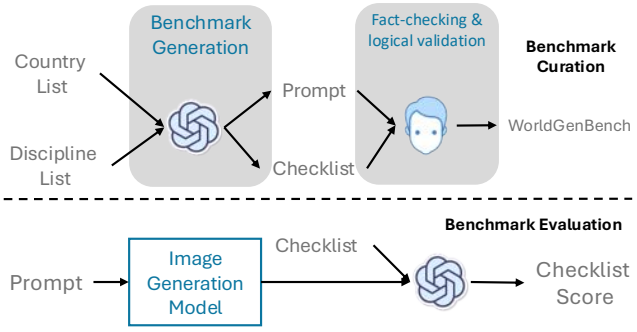


Figure 3. The Construction and Evaluation Pipeline of WorldGenBench.

3.1. WorldGenBench Construction

As shown in Figure 2 and Figure 3, our benchmark evaluates T2I models from two perspectives: Humanities and Nature. For the Humanities perspective, in order to reflect "world knowledge" and ensure fairness, we employ a large language model (LLM) to generate evaluation prompts covering 244 countries/regions worldwide, with three prompts per country, resulting in a total of 732 prompts related to history, culture, and related topics. For the Nature perspective, we similarly use an LLM to generate 340 evalu-

Benchmark	World Knowledge	Implicit Reasoning	Culture-awareness	Checklist Score
PhyBench[14]	✗	✓	✗	✗
Commonsense-T2I[6]	✓	✗	✗	✗
Wise[15]	✓	✗	✗	✗
WorldGenBench	✓	✓	✓	✓

Table 1. Comparison of T2I benchmarks across World Knowledge, Implicit Reasoning, Culture-awareness, and Checklist-based evaluation.

	Humanities							
Models \ Continent	AF	AN	AS	EU	NA	OC	SA	Avg.
FLUX.1-dev[11]	8.43	11.30	10.15	10.59	8.23	8.43	9.63	9.36
FLUX.1-schnell[11]	11.31	8.61	13.52	11.79	10.84	12.38	12.96	12.00
Playground-v2.5[12]	12.03	10.10	13.27	11.91	10.35	9.68	13.16	11.83
PixArt-alpha[1]	10.12	7.27	12.58	11.24	9.83	8.66	9.46	10.65
SDv3.5-Large[4]	11.82	13.24	13.43	12.72	11.46	11.91	15.57	12.57
SDv3.5-Medium[4]	12.08	11.94	12.44	11.40	10.15	13.33	12.69	11.85
SDXL[16]	10.89	9.09	11.47	10.14	9.94	9.54	10.74	10.55
HiDream-l1-Full*	16.61	13.61	17.96	18.28	14.39	16.53	13.79	16.68
Emu3[19]	10.44	9.35	11.85	12.57	9.84	10.00	11.77	11.13
JanusPro-1B[2, 13, 20]	3.34	5.93	4.02	2.97	2.07	3.60	5.33	3.41
JanusPro-7B[2, 13, 20]	6.87	5.45	8.57	9.22	5.28	5.24	8.46	7.41
JanusFlow-1.3B[2, 13, 20]	4.27	3.74	4.94	4.77	3.67	1.90	5.58	4.26
Show-o-512[23]	10.99	8.38	12.60	11.91	12.13	8.98	15.34	11.75
VILA-u-7B-256[22]	6.19	3.74	6.73	5.42	5.23	3.67	4.58	5.62
Harmon-1.5B[21]	10.04	7.69	10.86	9.56	8.99	9.05	12.17	9.96
Lumina-mGPT-2.0[17]	4.99	5.00	7.22	7.00	5.05	5.07	5.05	5.94
GoT-6B[5]	7.89	9.26	9.17	8.03	7.83	7.07	7.49	8.12
SimpleAR(SFT)[18]	7.97	8.28	8.40	7.28	7.61	6.63	8.21	7.75
SimpleAR(RL)[18]	7.40	4.75	8.13	8.91	7.13	7.61	8.92	7.90
Midjourney-v6†	11.62	9.01	13.21	14.16	11.34	10.23	12.05	12.33
Ideogram 2.0‡	10.65	5.93	14.77	13.37	11.35	12.84	10.42	12.42
GPT-4o§	23.96	17.47	26.22	25.12	24.98	21.03	23.29	**24.46**

Table 2. Knowledge Checklist Score on the Humanities Perspective: The first chunk corresponds to diffusion models, the second chunk to Auto-regressive models, and the third chunk to proprietary models. The results are organized according to continents. The corresponding full names are provided in the Appendix. 7.

ation prompts across 6 disciplines such as Astronomy and Physics. Subsequently, we conduct human verification to ensure the factual correctness and logical consistency of the benchmark prompts. Examples are shown in Figure 4 in the appendix.

3.2. Evaluation: Knowledge Checklist Score

Specifically, unlike all previous benchmarks, our benchmark does not directly evaluate image-text alignment, aesthetic quality, or related metrics. Instead, as shown in Figure 3, it focuses exclusively on assessing models' world knowledge and implicit reasoning capabilities. To this end, for each text-to-image prompt, we construct a corresponding checklist, where each item represents a specific attribute that we expect the T2I model to generate based on its internal knowledge and reasoning abilities. We then employ a state-of-the-art vision-language model, GPT-4o, to evaluate the generated images by determining the number of checklist items satisfied for each image. The Knowledge Checklist Score for an individual image is computed as the ratio of satisfied items to the total number of checklist items. The overall model performance is assessed by calculating the Average Knowledge Checklist Score across all generated images. The score is normalized to a range of 0 to 100.

*https://huggingface.co/HiDream-ai/HiDream-I1-Full

†https://www.midjourney.com/home

‡https://about.ideogram.ai/2.0

§https://openai.com/index/gpt-4o-system-card/

Models \ Fields	Nature						
	ASTR	BIO&MED	CHEM	EASC	PHYS	XDIS	Avg.
FLUX.1-dev[11]	5.15	4.08	4.33	5.41	7.50	3.02	5.19
FLUX.1-schnell[11]	8.39	5.12	4.83	7.50	9.20	5.67	6.87
Playground-v2.5[12]	5.66	3.27	1.23	4.27	2.33	2.27	3.07
PixArt-alpha[1]	5.75	4.57	1.55	3.46	2.09	2.86	3.19
SDv3.5-Large[4]	8.07	8.80	4.74	11.15	9.16	4.44	7.93
SDv3.5-Medium[4]	3.33	2.58	2.07	6.87	3.83	4.84	4.06
SDXL[16]	2.53	4.76	1.13	5.65	2.25	2.92	3.29
HiDream-l1-Full*	8.28	3.42	5.69	7.33	8.36	6.64	6.68
Emu3[19]	4.50	3.43	0.83	5.56	1.35	2.34	3.05
JanusPro-1B[2, 13, 20]	1.81	0.00	0.48	0.70	1.23	1.28	0.91
JanusPro-7B[2, 13, 20]	5.19	1.51	1.89	3.21	3.84	4.68	3.30
JanusFlow-1.3B[2, 13, 20]	0.57	0.00	1.67	0.34	0.43	0.94	0.60
Show-o-512[23]	7.01	1.48	2.78	3.39	5.03	3.34	3.76
VILA-u-7B-256[22]	1.56	1.25	1.15	4.39	3.28	1.84	2.46
Harmon-1.5B[21]	5.20	5.25	2.82	2.99	1.64	1.67	3.15
Lumina-mGPT-2.0[17]	2.03	0.64	1.11	2.12	1.73	0.60	1.41
GoT-6B[5]	2.62	0.69	1.57	1.36	1.33	1.96	1.53
SimpleAR(SFT)[18]	2.06	2.11	0.53	4.89	0.82	1.78	2.28
SimpleAR(RL)[18]	2.98	1.54	0.85	2.60	1.83	2.00	1.97
Midjourney-v6†	5.94	6.92	4.60	8.59	3.75	4.59	5.77
Ideogram 2.0‡	11.15	7.14	8.33	7.63	12.08	9.82	9.34
GPT-4o§	19.75	15.29	18.85	17.22	28.86	15.41	**19.61**

Table 3. Knowledge Checklist Score on the Nature Perspective: The first chunk corresponds to diffusion models, the second chunk to Auto-regressive models, and the third chunk to proprietary models. The results are organized according to continents. The corresponding full names are provided in the Appendix. 7.

4. Results

As shown in Table 2, we evaluated 22 state-of-the-art T2I models, including 8 advanced diffusion models, 10 auto-regressive models, and 3 proprietary models. We perform all evaluations using the default settings of each model.

Across both Table 2 (Humanities) and Table 3 (Nature), diffusion models remain the strongest open-source baseline: SD-v3.5-Large achieves the highest public scores with averages of 12.57 and 7.93, respectively. Within the autoregressive (AR) family, Show-o-512 leads its peers at 11.75 (Humanities) and 3.76 (Nature), confirming the promise of sequence-based generation for semantic coherence and local detail, yet still trailing the best diffusion model by roughly four points in scientific domains—evidence that AR methods must further improve world knowledge modeling and factual consistency. Proprietary systems outperform all open alternatives, with GPT-4o dominating at 24.46 (Humanities) and 19.61 (Nature), underscoring how extensive world knowledge and implicit reasoning confer robust cross-continental, cross-disciplinary generalization. Midjourney-v6 and Ideogram 2.0 reach diffusion-level performance in Humanities (12.33 and 12.42) but remain below ten points in Nature (5.77 and 9.34), indicating limited suitability for specialized scientific tasks. Based on this, although AR models demonstrate a high performance ceiling (as evidenced by GPT-4o), open-source AR models still lag significantly behind current diffusion models.

5. Conclusion

We introduced **WorldGenBench**, a benchmark designed to evaluate text-to-image models on world knowledge understanding and implicit reasoning. Through the proposed **Knowledge Checklist Score**, we provide a structured evaluation beyond surface-level text-image alignment. Experiments on 22 state-of-the-art models show that diffusion models remain strong among open-source systems, while proprietary models like GPT-4o demonstrate superior reasoning and knowledge integration. Our results highlight the need for future T2I models to move beyond pattern matching toward deeper understanding and inference.

References

[1] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α, 2023. 3, 4

[2] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 3, 4

[3] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835, 2021. 1

[4] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 3, 4

[5] Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, Xihui Liu, and Hongsheng Li. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. *arXiv preprint arXiv:2503.10639*, 2025. 3, 4

[6] Xingyu Fu, Muyu He, Yujie Lu, William Yang Wang, and Dan Roth. Commonsense-t2i challenge: Can text-to-image generation models understand commonsense? *arXiv preprint arXiv:2406.07546*, 2024. 2, 3, 1

[7] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023. 2

[8] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 2

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2

[10] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023. 2

[11] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 1, 3, 4

[12] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation, 2024. 3, 4

[13] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation, 2024. 3, 4

[14] Fanqing Meng, Wenqi Shao, Lixin Luo, Yahong Wang, Yiran Chen, Quanfeng Lu, Yue Yang, Tianshuo Yang, Kaipeng Zhang, Yu Qiao, et al. Phybench: A physical commonsense benchmark for evaluating text-to-image models. *arXiv preprint arXiv:2406.11802*, 2024. 2, 3, 1

[15] Yuwei Niu, Munan Ning, Mengren Zheng, Bin Lin, Peng Jin, Jiaqi Liao, Kunpeng Ning, Bin Zhu, and Li Yuan. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025. 2, 3, 1

[16] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 3, 4

[17] Alpha VLLM Team. Lumina-mgpt 2.0: Stand-alone autoregressive image modeling, 2025. 3, 4

[18] Junke Wang, Zhi Tian, Xun Wang, Xinyu Zhang, Weilin Huang, Zuxuan Wu, and Yu-Gang Jiang. Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl. *arXiv preprint arXiv:2504.11455*, 2025. 3, 4

[19] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 3, 4

[20] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024. 3, 4

[21] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation, 2025. 3, 4

[22] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. 3, 4

[23] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 3, 4

[24] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 1

[25] Daoan Zhang, Guangchen Lan, Dong-Jun Han, Wenlin Yao, Xiaoman Pan, Hongming Zhang, Mingxiao Li, Pengcheng Chen, Yu Dong, Christopher Brinton, et al. Seppo: Semi-policy preference optimization for diffusion alignment. *arXiv preprint arXiv:2410.05255*, 2024. 1

WorldGenBench: A World-Knowledge-Integrated Benchmark for Reasoning-Driven Text-to-Image Generation

Supplementary Material

6. Benchmark Comparison

Although recent text-to-image (T2I) benchmarks, such as PhyBench c[14], Commonsense-T2I [6], and Wise [15], have each explored specific abilities, they generally lack a systematic integration of world knowledge and implicit reasoning capabilities. PhyBench focuses on physical commonsense validation but lacks broad domain knowledge coverage; Commonsense-T2I incorporates basic world knowledge but offers limited depth in reasoning; and Wise advances knowledge application to some extent but remains insufficient in evaluating reasoning chains and factual inference. In contrast, WorldGenBench systematically integrates world knowledge comprehension and implicit reasoning as its core innovation, establishing an evaluation framework better aligned with the demands of complex, open-world tasks. World knowledge provides models with factual grounding across disciplines such as history, geography, and science, while implicit reasoning enables models to infer logically complete and contextually coherent scenes even from incomplete or underspecified prompts. This combination is crucial for future T2I systems: without world knowledge, generation often suffers from factual inaccuracies or distorted backgrounds; without reasoning ability, models struggle to handle the omissions, ambiguities, and implicit inferences ubiquitous in real-world text prompts. Thus, only by tightly coupling these two capabilities can T2I models achieve genuine understanding and faithful scene generation. Furthermore, WorldGenBench introduces culture-awareness as an auxiliary dimension to assess models' ability to adapt across diverse national and regional cultural contexts, thereby enhancing the benchmark's openness and diversity.

Moreover, WorldGenBench's introduction of the Knowledge Checklist Score represents a significant advancement over traditional VLM-based direct evaluation methods. Conventional VLM scoring approaches typically measure only the surface-level alignment between text and image, and are prone to misjudgments where superficial correlations mask deeper factual inconsistencies. In contrast, the Knowledge Checklist Score employs a pre-defined, fine-grained set of knowledge and reasoning checkpoints to rigorously verify whether key semantic elements are explicitly reflected in the generated image. This approach significantly reduces evaluation noise caused by VLM overgeneralization, associative inference, or linguistic bias. Each checklist item requires strict visual evidence and logical consistency, making the evaluation more objective, interpretable, and diagnostic in identifying specific deficiencies in a model's world knowledge and reasoning capabilities. This fine-grained, structured evaluation framework provides a necessary foundation for advancing T2I systems from surface-level pattern matching toward deep understanding and reasoning-driven generation.

6.1. WorldGenBench Example

7. Abbreviations

The abbreviations in the tables are:

Abbreviation	Continent Name
AF	Africa
AN	Antarctica
AS	Asia
EU	Europe
NA	North America
OC	Oceania
SA	South America
Avg.	Average

Table 4. Abbreviations and Full Names of Humanities.

Abbreviation	Discipline Name
ASTR	Astronomy
BIO&MED	Biology & Medicine
CHEM	Chemistry
EASC	Earth Sciences
PHYS	Physics
XDIS	Cross-Disciplinary
Avg.	Average

Table 5. Abbreviations and Full Names of Nature.

8. Evaluation Prompt for Knowledge Checklist Score,

We use GPT4o to make the evaluation.

The evalution prompt is:

Evaluation Procedure

1. For each checklist item, first read both the item and the explanation to fully understand the complete semantic requirement.

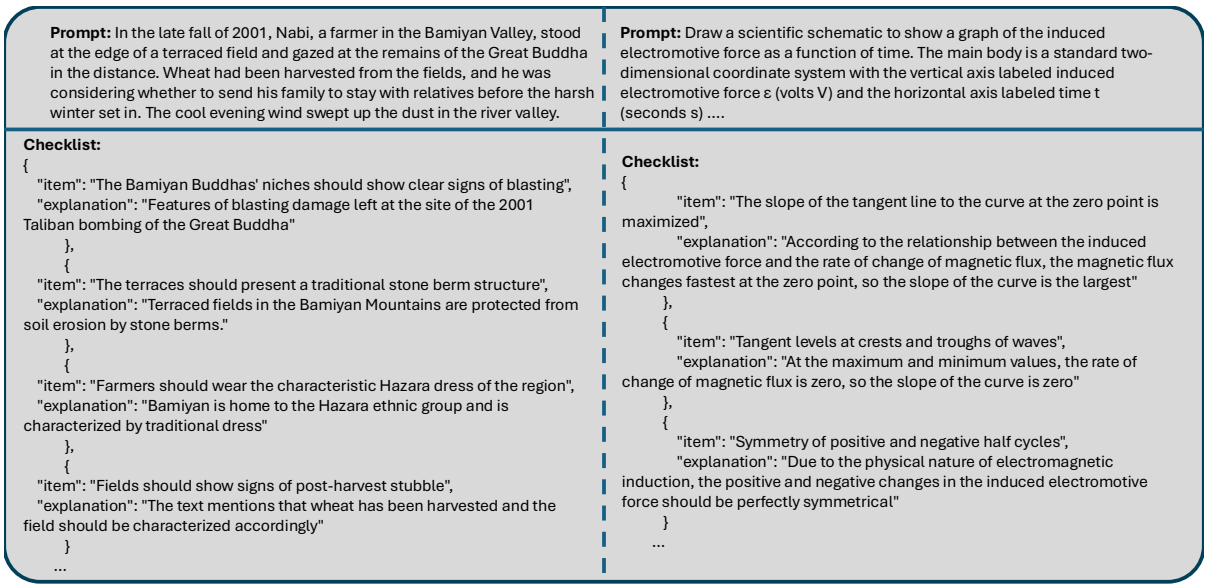


Figure 4. Examples from WorldGenBench: Input Prompt and Corresponding Checklist. Left: Humanities; Right: Nature.

2. Then examine the image and determine whether it explicitly, fully, and unambiguously shows all visual elements required by the semantic meaning of the item.
3. Apply strict criteria: if any required element is missing or unclear, you must judge the item as Not Satisfied (0), even if some parts are present.
4. Do not infer or assume meanings based on keyword similarity, visual resemblance, symmetry, or general scientific knowledge. You may only use what is explicitly shown or labeled in the image.
5. Any unmarked, ambiguous, inferred, or implied content must be treated as not provided.

Strict Interpretation Rules

- The presence of a label or term (e.g., "Standard") does not imply satisfaction of a requirement (e.g., "relative position of standard pressure") unless the actual visual structure is present (e.g., scale, baseline, reference line).
- Arrows, colors, directions, or shapes must be explicitly defined in the diagram or legend (e.g., as representing force, pressure, volume) to count as valid evidence.
- A checklist item is only satisfied if its entire explanation is visually and explicitly fulfilled. If there is even one missing or ambiguous component, return 0.

Reverse Verification Requirement After completing the initial pass, re-check all items judged as "1 (Satisfied)" by asking:

- Is there any required detail that was not explicitly shown or labeled?
- Was the decision made based on assumption, familiarity, or similarity, instead of strict visual evidence?

If yes, correct the judgment to 0 (Not Satisfied).

Example

- **Item:** "The volume change should be indicated with an upward or downward arrow labeled 'Volume'."
- **Explanation:** "This shows that the diagram must make the direction and meaning of volume change visually explicit."
- $\rightarrow$ If the diagram contains an arrow labeled 'Volume' clearly pointing up or down, return 1.
- $\rightarrow$ If there is just an arrow without label, or just the word 'Volume' without direction, return 0.

Output Format Return a list of N binary values (0 or 1), where each value corresponds to the same-positioned checklist item:

- 1 = Fully satisfied based on explicit visual evidence
- 0 = Not satisfied due to missing, ambiguous, or incomplete visual evidence

Example output for 3 checklist items:

$$[1, 0, 0]$$

Image {`image`}

Checklist {`checklist`}

Please evaluate the image strictly following the above procedure and directly return the binary list.

9. Visual Cases

We present two cases to compare models and one case to demonstrate the model's success and failure.

The cases for model comparison are shown in Figure 5 and Figure 6, while the case illustrating the detailed checklist scores is presented in Figure 7.

Prompt: In the late summer of 2001, a monk at the Echmiadzin Monastery was organizing medieval manuscripts in the ancient library. Sunlight streaming in through the stained-glass windows illuminates the exquisite Armenian miniatures on parchment. In the distance, the sounds of traditional hymn practice are heard, and nothing here seems to have changed over the millennia.

Figure 5. Visual Case 1: An example case from the humanities domain, including the prompt and the results from GPT-4o, HiDreaml1-Full, and Show-o. Checklist is in 9.1

9.1. Checklist for Visual Case 1

Item: The picture should feature typical Armenian church architecture
Explanation: The Echmiadzin Monastery has a unique architectural style, including conical domes and stone decorations.

Item: Traditional style wooden bookshelves should appear
Explanation: As an ancient monastic library, it has retained its historic and traditional furnishings.

Item: The friar shall wear the black traditional robe
Explanation: The monks of the Armenian Apostolic Church have their own specific traditional dress code.

Item: There should be specialized tools on the desktop for restoration
Explanation: Antiquarian book restoration requires specific specialized tools, which are necessary working instruments.

Item: Windows should present a characteristic painted pattern
Explanation: Stained glass in churches often contains religious themes and traditional motifs.

Item: Manuscripts should have typical Armenian text
Explanation: Ancient Armenian manuscripts use their unique alphabet system.

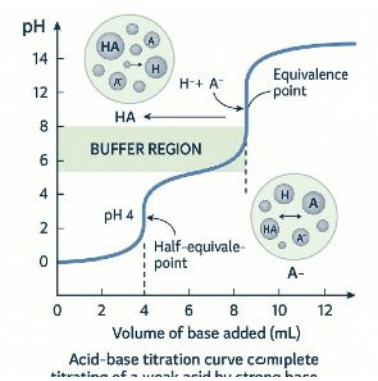
Item: There should be incense burners and candles in the room
Explanation: These are the traditional liturgical items of the Armenian Church, which are used year-round.

Item: The walls should have old frescoes
Explanation: Monastery interiors often preserve historic religious frescoes.

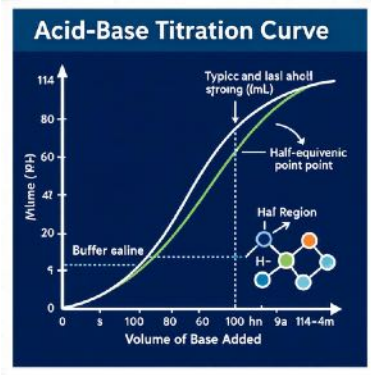
Item: The light should create a specific projection
Explanation: The text mentions sunlight filtering through the colored windows, a light effect that is one of the features of the church building.

Item: Stone floors should be featured
Explanation: Monastery buildings are paved with local stone, reflecting architectural tradition.

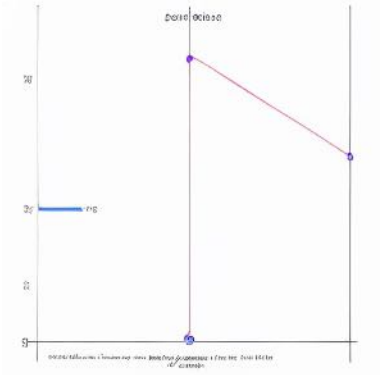
Prompt: Create a detailed pedagogical illustration of an acid-base titration curve showing the complete titration of a weak acid by a strong base. The graph contains a large XY coordinate system with the X-axis representing the volume of base added (mL) and the Y-axis representing the pH (0-14). The curve should show a typical S-shape with the abrupt jump points clearly visible. Use a different color to highlight at the buffer region (approximately pH 4-6) and mark it with an arrow. Equivalent and half-equivalent point locations need to be labeled, and specific pH values should be labeled at key points. Diagrams should contain illustrations at the molecular level showing the dynamic equilibrium process of HA/A-. Use a professional scientific graphical style with clear illustrations. The overall color scheme should be professional and easy to read, blue or green is recommended.



GPT4o: 0.7



Hidreaml1: 0.4



Show-o: 0.0

Figure 6. Visual Case 2: An example case from the Nature domain, including the prompt and the results from GPT-4o, HiDreaml1-Full, and Show-o. Checklist is in 9.2

9.2. Checklist for Visual Case 2

Item: The curve shows a smoother slope in the buffer region
Explanation: The buffer zone is resistant to pH changes, so the curve changes slowly in this region.

Item: The curve shows a steep change near the equivalence point
Explanation: When the equivalence point is reached, the system is extremely sensitive to pH changes, and small additions of reagents can lead to dramatic pH changes.

Item: The pH at the half-equivalent point should be equal to the pKa of the weak acid
Explanation: According to the Henderson-Hasselbalch equation, when [HA] = [A-], pH = pKa.

Item: There is a significant difference between the slopes of the curves on either side of the buffer region
Explanation: Reflecting the difference in sensitivity of the system to pH changes at different stages.

Item: Proportion of HA and A^- in molecular level insets versus curve
Explanation: The concentration of acid ions gradually increases and the concentration of un-ionized acid molecules decreases during the titration process.

Item: Coordinate axis scales are evenly spaced and clear
Explanation: Specialized scientific charts require accurate data representation.

Item: Buffer labeling should point to the inflection area of the curve
Explanation: The region of most significant buffering effect is near the half-equivalent point.

Item: Legends contain math formulas or chemical equations
Explanation: Professional titration curves usually need to show the relevant theoretical basis.

Item: The pH value at the equivalent point should be greater than 7
Explanation: Weak acids react with strong bases and show basicity at the equivalence point.

Item: The pH at the start of the curve should be close to

the pH of the weak acid
Explanation: The initial state of the titration reflects the acidic character of the original solution.

Prompt: In December 1982, deep in the rainforest of the Moumba province in western Gabon, a honey collector from the Miéné tribe was engaged in the traditional collection of wild honey. It's the end of the dry season, when local wild bee colonies are at their most active. He skillfully navigates his way through the tall canopy layers, using techniques passed down from his ancestors.

GPT4o: 0.6

Part of the checklist:

{
"item": "The honey gatherer wore simple shorts and was topless",
"explanation": "Conforms to the tropical jungle environment and traditional honey harvesting operations, and is easy to climb."
},
{
"item": "A tool belt woven from vines was tied around his waist.",
"explanation": "Carrying traditional tools necessary for honey harvesting, demonstrating the utilization of indigenous materials"
},
{
"item": "Hand-held traditional smoking tools made from species-specific wood",
"explanation": "Necessary tools to calm the colony, reflecting traditional techniques"
},
{
"item": "The surrounding trees are typical of Gabon's western rainforest vegetation, such as African pearwood.",
"explanation": "Specialty tree species in Muamba province, providing wild bee colony habitats"
},
{
"item": "The scene has typical rainforest vines",
"explanation": "Local tropical rainforest ecosystem characteristics that provide climbing aids for honey harvesting"
},
{
"item": "Bark ropes for protection were wrapped around his feet.",
"explanation": "Safety measures for traditional tree climbing techniques reflecting indigenous knowledge"
}
...

Figure 7. Visual Case 3: An example case from GPT4o. In the checklist, red text indicates that the model did not score on the corresponding item, while green text indicates that the model received a score.