Comprehensive Evaluation Frameworks and Future Directions: Current State and Challenges of Text-to-Image Generation Models

Linpeng Shang*

Department of Electrical and Computer Engineering, Colorado State University, Colorado, USA

*Corresponding author: magicstar6966@gmail.com

Abstract. Text-to-image (T2I) generation models have great innovations in generative AI. However, evaluating these models could still be regarded as a crucial challenge. Traditional metrics include Frechet Inception Distance (FID) and CLIPScore, and they concentrate on image quality and text-image alignments. However, they often fail to address fine-grained details, for example, spatial relationships and attribute binding. Recent frameworks, for example, GENEVAL, TIFA, HEIM, and HRS-Bench, have involved multidimensional evaluation methods. They have combined automated metrics and human preference-based approaches. Although these frameworks improve interpretability and comprehensiveness, they still face limitations. To be more specific, they are not good enough in scalability or standardization. Moreover, they lack emphasis on social and ethical considerations. These methods often do not have enough granularity. This paper will analyze the advantages and disadvantages of the current evaluation methods, and highlight the needs for modular, interpretable, and scalable systems. They integrate automated metrics with selective human feedback. By dealing with these challenges, T2I models can better cater to people's expectations. In addition, they support various applications in art and data generation. This study would offer practical perspectives for promoting progress in the evaluation and development of next-generation T2I models.

Keywords: Text-to-Image Generation; Performance Evaluation; Human Preferences; Comprehensive Benchmarks; Generative AI.

1. Introduction

Recently, the high-speed development of T2I models has given rise to huge changes in the field of generative models. Since the appearance of models like Stable Diffusion, these new generative models have gained close attention. They are based on diffusion techniques and multimodal pretraining [1]. They have profound progress in visual generation [1]. These models showed huge potential in applications such as art design, and data generation. However, as model generation capabilities improve, how to assess them with accuracy has become a challenge in the current research.

Traditional evaluation metrics are effective in evaluating image quality and text-image alignment. However, they might not get more complex details in generation tasks. For instance, for spatial relationships between multiple objects and attribute binding, they might not done well in exploring details [2]. For instance, CLIPScore assesses alignment by assessing the similarity between image and text embeddings. However, these metrics might not perform well when dealing with complex relationship reasoning tasks. To deal with these limitations, some more detailed and various evaluation frameworks have appeared, such as GENEVAL and TIFA [3,4].

The GENEVAL framework could use object detection models to evaluate the co-occurrence and color attributes of objects, in the generated image [3]. This has quantified the model's fine-grained expression ability of the text. TIFA, on the other hand, could evaluate text-image fidelity. It uses Visual Question Answering (VQA). It proposed a reference-free evaluation way on the basis of automatically generated questions and answers [4]. This method has provided interpretable evaluation results in multiple dimensions.

Besides automated evaluation methods, those evaluation approaches based on human preference are also crucial in assessing T2I models. For example, Human Preference Score (HPS) proposed by

Wu et al. and ImageReward proposed by Xu et al., model human preferences to enhancing the semantic alignment and generation quality of the models [5, 6]. Particularly, ImageReward was built from over 130,000 expert annotations [6]. It could effectively get the aesthetic and semantic features of generated images [6]. In this way, it can be said to be a profound automated evaluation metric.

Current problems include the single-dimensional evaluation and the lacking of variety in existing evaluation methods. Under the circumstance, comprehensive benchmark frameworks such as HEIM and HRS-Bench have appeared. HEIM evaluates models on 12 key dimensions which include aesthetics, reasoning, and multilingual abilities [7]. It could reveal the limitations of current models in the face of different generation tasks [7]. HRS-Bench could evaluate models based on their performance in five main categories, including accuracy, robustness, generalization, fairness, and bias [8]. They across 13 skills, for which it could offer a comprehensive and specific assessment [8]. It cannot be denied that generative models do well in some aspects. However, they still face huge challenges in dealing with complex scenarios and tasks which contain social and ethical considerations.

There is progress in combining automated and human evaluation methods. However, current evaluation frameworks still face challenges in the relationship between fine-grained analysis and scalability. Automated evaluations have obvious advantages in efficiency. But they do not have sufficient interpretability. Human evaluations are reliable, but they are costly. Therefore, future research should try to develop interpretable evaluation metrics. They can be integrated with automated systems and provide a more comprehensive assessment of models. Standardized evaluation protocols and dynamic human feedback mechanisms will also be crucial. After all, they allow fair comparison and continuous improvement of T2I generation models.

This paper aims to analyze these challenges and advancements. It also analyzes the current state of evaluation for T2I models. The next chapters will focus on the limitations of existing benchmarks, methods for aligning models with human preferences, and comprehensive evaluation frameworks. Through an integrated analysis of these various methods, this study will provide effective perspectives for the evaluation and development of the next generation of T2I generation models.

2. Related Works

In the T2I generation, the evaluation and benchmarking of models have gained more and more attention. It is because of the rapid progress in generative techniques such as diffusion models and multimodal pretraining. Existing studies have identified some main challenges. These challenges include the alignment of generated images with textual prompts and capturing fine-grained attributes, as well as the highlights on human preferences. This section has done literature review on evaluation frameworks, human-centric approaches, and holistic benchmarking methodologies.

2.1.Evaluation Frameworks.

Traditional evaluation metrics are widely used to measure image quality and text-image alignment. For example, FID and CLIPScore. However, these metrics might not capture compositional and fine-grained capabilities of T2I models [9, 2]. Ghosh et al. introduced GENEVAL which is an object-focused evaluation framework. It leverages object detection models to evaluate attributes like object co-occurrence, spatial relationships, and color accuracy [3]. It achieves a strong alignment with human judgments and also reveals limitations in tasks such as spatial reasoning and attribute binding [3].

Moreover, efforts have been made to extend evaluation frameworks to include wider dimensions. TIFA has involved a question-answering-based method to evaluate text-image faithfulness. It can offer interpretable and fine-grained assessments across different tasks [4].

2.2.Human-Centric Approaches.

Human preferences are important in assessing the quality of T2I models. Wu et al. proposed the HPS. It could reveal important gaps in authenticity and compositionality compared to human expectations [5]. To address these gaps, a framework was developed to increase the semantic alignment and aesthetic quality in generated images [6]. It is ImageReward, a reinforcement learning framework with the use of over 137,000 annotations [6]. Moreover, SeeTRUE used a question-production and answering way to localize specific text-image misalignments. It outperforms prior methods in identifying semantic inconsistencies [10].

2.3.Holistic Benchmarks.

Comprehensive evaluation frameworks have been developed to capture the various nature of T2I models. Lee et al. introduced HEIM. It is a benchmark assessing 12 critical dimensions, including aesthetics, reasoning, and so on. Their results showed a lack of a single model excelling in all dimensions. Moreover, they emphasized the need for standardization in evaluation protocols [7]. Similarly, Petsiuk et al. provided a multi-task benchmark with varying difficulty levels. It has demonstrated that even state-of-the-art models like DALL-E 2 and Stable Diffusion could face huge challenges in dealing with complex prompts [11].

3. Holistic and Human-Aligned Evaluation Frameworks

Evaluating T2I generation models have focused on a narrow range of metrics in the past. They concentrated mainly on generic image quality and text-image correspondence. However, these models are gradually making progress and begin to use more and more complex and various applications. In this way, conventional benchmarks and evaluation methodologies have revealed important limitations. Beyond basic image fidelity and alignment, more nuanced aspects must be comprehensively assessed, which include compositionality, social and ethical considerations, and long-term adaptability. Recent efforts reflect a move towards more holistic and human-aligned evaluation frameworks. They contain object-level metrics, scenario diversity, human preferences, and broad skill dimensions. These new ways are integrations of automated and human judgment. They highlight key gaps in model performance and lay a foundation for future solutions. Those solutions might integrate scalable metrics and social responsibility. The sections below have surveyed these benchmarks, and also analyzed their advantages and disadvantages. Moreover, potential directions, for the purpose to develop more comprehensive and equitable evaluation standards for T2I generation models, will be discussed in the following sections.

3.1.Limitations of Current Benchmark Evaluation Dimensions

The evaluation benchmarks concentrated on limited dimensions in the past, for which they have some shortcomings. Metrics like FID are commonly used. However, they might not be able to capture the finer-grained capabilities of the models [9].

To deal with these shortcomings, GENEVAL has made use of an object-oriented framework. It evaluates attributes such as object count, and spatial relationships [3]. Its high correlation with human evaluations could prove that it is effective [3]. Also, it reveals great gaps in tasks such as spatial reasoning [3]. Similarly, HRS-Bench expanded the evaluation scope by assessing performance across 13 skill categories in 50 diverse scenarios [8]. These can better overcome the shortcomings of past metrics.

GENEVAL excels in granular analysis by identifying compositional errors, whereas HRS-Bench offers broader coverage in social and ethical dimensions [3, 8]. Another comprehensive benchmark, HEIM, evaluates 12 aspects, including aesthetics, multilingual capabilities, and toxicity, combining automated and human evaluations to enhance reliability [7]. However, for medium to high-difficulty tasks, HRS-Bench revealed that most models fail to generate effective results, such as in complex spatial relationships and action compositions [8]. Additionally, GENEVAL struggles with visually

unrealistic images, and HEIM's reliance on human evaluations makes it costly and difficult to scale [3, 7].

Future evaluation frameworks need to integrate fine-grained yet scalable metrics for a comprehensive understanding of model performance. Emphasis should be placed on combining modular automated metrics with selective human verification to address nuanced aspects such as compositional reasoning and aesthetic appeal. Furthermore, standardized evaluation protocols will enhance reproducibility and fair comparisons between models.

3.2. Misalignment Between Generated Images and Human Preferences

T2I generation models often fail to meet human preferences in terms of aesthetics and composition. For quantifying human-centric alignment, the HPS was introduced. It could reveal huge gaps in realism and compositionality [5]. Based on this, ImageReward leveraged over 137,000 annotated comparisons and improved semantic alignment and aesthetic quality through Reinforcement Learning with Human Feedback (RLHF) [6]. It could reflect limitations in earlier methods like CLIPScore [6]. Moreover, SeeTRUE evaluated and pinpointed misalignments between textual prompts and generated images through a Q&A pipeline [10]. It outperforms baseline models on datasets like Winoground and DrawBench [10].

While HPS provides a measure of consistency with human preferences, the dataset's limitations may cause scores to reflect merely a small subset of human preferences [5]. Similarly, ImageReward faces challenges in dataset diversity. Although it uses the large-scale DiffusionDB prompt dataset, its annotations are constrained by resource limitations [6]. It could not represent users' practical needs. SeeTRUE covers various datasets, including real and synthetic ones [10]. However, it suffers from distributional discrepancies with real-world data, which could limit the model's generalizability.

Although HPS, ImageReward, and SeeTRUE have greatly improved human preference alignment and semantic evaluation accuracy, they can still be improved. Future efforts should be made to collect more various datasets. The data should cover different cultural contexts and parts to increase model generalizability. From the technical perspective, integrating multimodal pre-trained models and visual Q&A methods (e.g., VQ2) could better improve complex semantic alignment between text and images [10]. Additionally, research should highlight the safety and fairness of generated content. At the same time, research should avoid potential biases or psychological harm. Moreover, studies should provide transparent, explainable evaluation mechanisms. Through these improvements, T2I generation models could better meet practical needs.

3.3.Holistic Evaluation Frameworks

In recent years, researchers have put more and more focus on multifaceted holistic evaluation frameworks. HRS-Bench can be said to be a good example. It evaluates T2I models across 13 skills and 50 scenarios, which is comprehensive [8]. Meanwhile, it has made use of new alignment evaluation metrics [8]. HEIM makes evaluation on models across 62 scenarios, which is systematic [7]. It adds dimensions including originality and ethical risks [7]. Petsiuk et al. proposed a multitask benchmark. Through task difficulty levels, it could capture model adaptability and robustness [11].

Both HRS-Bench and HEIM mentioned before highlighting wide evaluations, but their methods differ from each other. HRS-Bench focuses on automated evaluations for scalability [8]. Differently, HEIM could integrate human judgment for nuanced tasks like aesthetics capabilities [7]. However, neither fully deals with time dimensions. For instance, the impact of fine-tuning or long-term usage on performance. Petsiuk et al.'s multitask benchmark introduced task-specific adaptability. Although it can handle images involving sensitive attributes like gender or race [11]. However, it does not have enough consideration for broader social issues such as fairness and bias mitigation [11].

Future frameworks should adopt modular architecture. It could evaluate both technical performance and social influence. Integrating longitudinal research in benchmarks could reveal the ways in which models adapts to evolving use cases and fine-tuning practices. Moreover, containing ethical evaluation standards will be key to the development of socially responsible models.

4. Outlook and Challenges

Although T2I generation models have made significant progress in recent years, achieving comprehensive and sustainable performance improvements still faces several challenges. As generative technologies continue to evolve, evaluation methods must also innovate to handle increasingly complex and diverse generation tasks. Future research should focus on the following areas:

4.1.Integration of Fine-grained and Multidimensional Evaluation

Current evaluation methods tend to focus on one dimension or fine-grained analysis, such as spatial relationships, object counting, or attribute alignment. However, the complexity of generation tasks extends far beyond these dimensions. Future evaluation frameworks should integrate multiple dimensions, such as aesthetics, reasoning ability, bias and fairness, and ethical concerns. While frameworks like GENEVAL and TIFA have made initial progress, combining fine-grained analysis with comprehensive multidimensional evaluation to ensure a holistic assessment of model performance remains an important research challenge.

4.2.Deep Integration of Automation and Human Feedback Mechanisms

Although current automated evaluation systems are efficient, they often lack sufficient interpretability and struggle with complex T2I alignment tasks. Human evaluation, though reliable, is costly and labor-intensive. Future research should aim to develop modular and interpretable automated evaluation systems that can dynamically integrate human feedback, enabling continuous improvement and optimization of model performance. The challenge will be to balance high-quality evaluations with improved efficiency.

4.3.Integration of Social Ethics and Fairness Issues

With the widespread use of T2I generation technology, ensuring that generated content is free from bias, discrimination, or harmful information has become a critical social and ethical issue. Frameworks like HRS-Bench have started exploring model fairness and robustness, but how to systematically integrate these social and ethical considerations into evaluation frameworks, especially in multilingual and multicultural contexts, remains an open question. Addressing these issues will contribute to making technology socially responsible and reduce potential risks in real-world applications.

4.4.Establishment of Standardized Evaluation Protocols

Currently, different studies use varying evaluation protocols, datasets, and standards, making direct comparisons between models difficult. Future research should promote the establishment of standardized evaluation protocols. It will provide a unified benchmark for academic research and industry applications of T2I generation models. Standardized evaluation protocols will also promote cross-disciplinary collaboration. They could drive the adoption of generative models in more application scenarios as well.

5. Conclusion

This paper analyzes the current state and future trends in the evaluation of T2I generation models. As generative models like Stable Diffusion keep evolving, the need for robust and multi-dimensional evaluation frameworks has become more and more evident. Traditional metrics include FID and CLIPScore. They are foundational in evaluating basic image quality and text-image alignment. Nevertheless, they are not enough in capturing the complex aspects of modern T2I models. These limitations have hindered the development of more sophisticated evaluation frameworks. For instance,

GENEVAL and TIFA. They offer finer-grained assessments of object co-occurrence, spatial relationships, and text fidelity, so they could be more corresponding to human judgment.

Human-centric evaluation approaches have further improved the evaluation landscape by containing human aesthetic and semantic preferences. These methods bridge the gap between automated metrics and human perception. So, it could ensure that generated images meet subjective quality standards.

Although there are huge advancements, a lot of challenges still exist. The challenges in the relationship between fine-grained analysis and scalability persist. Automated methods have a lack of interpretability. Moreover, human evaluations are resource-intensive and integrating social and ethical considerations into evaluation frameworks is also a challenge. It is more obvious in multicultural and multilingual contexts. The need for standardized evaluation protocols is key to promoting fair comparisons and reproducibility across studies. The development of efficient multi-task learning frameworks is crucial to deal with complex generation tasks.

Future research should put priority on the integration of fine-grained and multidimensional evaluation metrics. Also, the deep incorporation of human feedback into automated systems deserves close attention. Moreover, the systematic inclusion of social ethics and fairness issues within evaluation frameworks should be focused on as well. In the face of these challenges, the assessment of T2I generation models will be more reliable. Finally, it will promote models' development.

To conclude, although huge developments have been made in evaluating T2I, more innovations in evaluation methods are still required. It is crucial for the methods to keep pace with the development speed of generative technologies. Efforts should be made to develop integrated and ethical evaluation frameworks. It can make sure that T2I generation models remain reliable and their social responsibility. It can also expand their potential in various applications.

References

[1] Tang R, Liu L, Pandey A, Jiang Z, Yang G, Kumar K, et al. What the daam: Interpreting stable diffusion using cross attention. arXiv preprint arXiv:2210.04885. 2022.

[2] Hessel J, Holtzman A, Forbes M, Bras R L, & Choi Y. Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718. 2021.

[3] Ghosh D, Hajishirzi H, Schmidt L. GENEVAL: An object-focused framework for evaluating text-to-image alignment. Proceedings of NeurIPS. 2023.

[4] Hu Y, Liu B, Kasai J, Wang Y, Ostendorf M, Krishna R, Smith N A. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 20406-20417). 2023.

[5] Wu H. Human Preference Score: Better aligning text-to-image models with human preference. Proceedings of ICCV. 2023.

[6] Xu J, Liu X, Wu Y, Tong Y, Li Q, Ding M, Tang J, Dong Y. ImageReward: Learning and evaluating human preferences for text-to-image generation. Proceedings of NeurIPS. 2023.

[7] Lee T, Yasunaga M, Meng C, Mai Y, Park JS, Gupta A, et al. Holistic Evaluation of Text-to-Image Models (HEIM). Proceedings of NeurIPS. 2023.

[8] Bakr E M, Sun P, Shen X, Khan F F, Li L E, Elhoseiny M. HRS-Bench: Holistic, reliable and scalable benchmark for text-to-image models. Proceedings of ICCV. 2023.

[9] Jayasumana S, Ramalingam S, Veit A, Glasner D, Chakrabarti A, Kumar S. Rethinking fid: Towards a better evaluation metric for image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9307-9315). 2024.

[10] Yarom M, Bitton Y, Changpinyo S, Aharoni R, Herzig J, Lang O, Ofek E, Szpektor I. What you see is what you read? Improving text-image alignment evaluation. Proceedings of NeurIPS. 2023.

[11] Petsiuk V, Siemenn A, Surbehera S, Chin Z, Tyser K, Hunter G, et al. Human evaluation of text-to-image models on a multi-task benchmark. NeurIPS 2022 Workshop on Human Evaluation of Generative Models. 2022.