Re-Thinking the Automatic Evaluation of Image-Text Alignment in Text-to-Image Models

Anonymous ACL submission

Abstract

Text-to-image models often struggle to generate images that precisely match textual prompts. Prior research has extensively studied the evaluation of image-text alignment in text-to-image generation. However, existing evaluations primarily focus on agreement with human assessments, neglecting other critical properties of a trustworthy evaluation framework. In this work, we first identify two key aspects that a reliable evaluation should address. We then empirically demonstrate that current mainstream evaluation frameworks fail to fully satisfy these properties across a diverse range of metrics and models. Finally, we propose recommendations for improving image-text alignment evaluation.

1 Introduction

Text-to-Image (T2I) models have demonstrated remarkable capabilities in generating high-quality, realistic images (Betker et al., 2023; Esser et al., 2024). Despite these advances, they still face challenges in accurately interpreting and adhering to user prompts. Common failures include generating incorrect object counts, attributes, or spatial relationships (Li et al., 2024). Nevertheless, evaluating text-image alignment remains a persistent and unresolved problem in the field.

Several frameworks exist for evaluating image-text alignment. Model-based approaches include CLIPScore (Hessel et al., 2021) and VQAScore (Lin et al., 2024). Component-based methods decompose text prompts into fine-grained elements and assess alignment through techniques like question generation and answering (QG/A) (Hu et al., 2023; Cho et al., 2023). Additionally, detector-based frameworks such as UniDet-Eval (Huang et al., 2023) leverage object detection for evaluation.

Despite the vast number of evaluation frameworks, few studies have thoroughly investigated the trustworthiness of image-text alignment assessment. Most existing work focuses narrowly on aligning automatic evaluation results with human judgments (Li et al., 2024; Wiles et al., 2025). When addressing trustworthiness, researchers primarily concentrate on improving human evaluation protocols to better validate automatic metrics (Otani et al., 2023; Wiles et al., 2025). However, these approaches overlook the critical point that a truly trustworthy evaluation framework must encompass more dimensions than mere correlation with human assessment. Additional discussion of related works is provided in Appendix A.

Figure 1: A robustness failure case of VQAScore. The right image is visually the same as the left one, yet their calculated VQAScore differs a lot.

In this work, we identify two critical properties for trustworthy evaluation frameworks: **Robustness** and **Significance**. Through empirical analysis, we demonstrate that current evaluation methods fail to fully satisfy these criteria, highlighting an important research direction for improving text-image alignment assessment. For example, as illustrated in Figure 1, while two visually similar images receive substantially different VQAScore evaluations, revealing clear deficiencies in Robustness.

In summary, our work makes two key contributions: (1) identifying Robustness and Significance as fundamental requirements for trustworthy evaluation frameworks, and (2) systematically demon-

strating how current frameworks fail to meet these criteria.

2 Methodology

2.1 Preliminaries

We would like to briefly introduce the common image-text alignment evaluation framework that we discuss in this research. Generally, given a benchmark consisting a set of N prompts $P = \{p_1, ..., p_N\}$, a text-to-image model M is used to generate images $I = \{I_1, ..., I_N\}$ based on these prompts. [1] An automatic metric J is used to evaluate the alignment between the image-text pair (p_i, I_i), and outputs a score $s_i = J(p_i, I_i)$. The final evaluation result is the average of all scores: $s_M = \frac{1}{N}\sum_{i=1}^{N} s_i$. If there are multiple models $M_1, ..., M_K$, the scores $s_{M_1}, ..., s_{M_K}$ also provides a ranking of these models, which is also a part of the evaluation result that people focus on.

2.2 Aspects

We would like to propose two important aspects that a trustworthy image-text alignment evaluation should focus on.

Robustness Robustness requires that evaluation results remain consistent under reasonable perturbations of the input pair (p_i, I_i). In this work, we specifically examine two critical dimensions of robustness: (1) robustness to randomness and (2) robustness to image perturbations.

Robustness to Randomness Most state-of-the-art text-to-image models employ denoising diffusion processes, where the generation output for a given prompt p_i depends on the sampled noise prior. This inherent randomness introduces variability in evaluation results. A trustworthy evaluation framework must maintain consistent model rankings despite this randomness - otherwise, the evaluation fails to reliably indicate which model performs better. To assess robustness under randomness, we systematically evaluate model performance across different random seeds.

Robustness to Image Perturbation We make a fundamental assumption that for visually similar images I_i and $I_i^{'}$, their evaluation scores $J(p_i, I_i)$ and $J(p_i, I_i^{'})$ should also be close. Large discrepancies would indicate potential metric flaw rather than true model capability. To assess robustness against image perturbations, we apply a minimal transformation: given an image I with pixel values $I_{c,h,w} \in [0, 255]$, c, h, w corresponds to the channels, height and width of the image, we achieve a perturbed image $I^{'}$ with pixel values [2]:

$$I^{'}_{c,h,w} = \begin{cases} I_{c,h,w} + 1 & \text{if } I_{c,h,w} < 255 \\ I_{c,h,w} & \text{otherwise} \end{cases} \quad (1)$$

The robustness metric is computed by calculating the performance gap as:

$$\Delta J_i = |J(p_i, I_i) - J(p_i, I_i^{'})| \quad (2)$$

Significance The property of Significance examines whether an observed performance difference (e.g., $s_{M_1} > s_{M_2}$) reflects a meaningful superiority of model M_1 over M_2. To quantify this, we employ two complementary approaches:

1. Statistical Testing: For evaluation score sets $S_{M_1} = (s^1_{M_1}, ..., s^N_{M_1})$ and $S_{M_2} = (s^1_{M_2}, ..., s^N_{M_2})$, we conduct a paired t-test to assess statistical significance.

2. Dominance Ratio: We compute the empirical probability of M_1 over M_2 as:

$$R = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[s^i_{M_1} > s^i_{M_2}] \quad (3)$$

3 Experiment Setup

For our evaluation, we select four widely-used text-to-image generation models that represent diverse architectures and capabilities: Stable-Diffusion-3 (SD3) (Esser et al., 2024), Stable-Diffusion-XL (SD-XL) (Podell et al., 2023), Stable-Diffusion-1.5 (SD1.5) (Rombach et al., 2022), and PixArt-Sigma-XL (Pixart) (Chen et al., 2024a). This carefully chosen set ensures comprehensive coverage of the current state-of-the-art in diffusion-based text-to-image generation.

For metrics, we employ three widely applicable metrics: CLIPScore (Hessel et al., 2021) , VQAScore (Li et al., 2024) , and DSGScore (Cho et al., 2023). For benchmarking, we use MSCOCO (Lin

[1] Though there can be multiple images generated using a certain prompt, many practices only generate one image per prompt.

[2] We confirm the perturbation's visual imperceptibility through manual inspection of multiple cases, verifying that modified images remain indistinguishable from originals.

Model Name	VQAScore			CLIPScore			DSGScore		
	42	3407	5096	42	3407	5096	42	3407	5096
Stable-Diffusion-3	91.18(1)	90.90(1)	91.06(1)	26.39(1)	26.34(1)	26.36(1)	93.66(1)	91.99(1)	93.52(1)
Stable-Diffusion-XL	86.63(3)	86.01(3)	85.77(3)	25.93(2)	25.81(2)	25.85(2)	89.68(3)	90.04(3)	90.44(2)
Pixart-Sigma-XL	87.04(2)	87.16(2)	86.72(2)	25.78(3)	25.71(3)	25.75(4)	90.52(2)	90.53(2)	89.99(3)
Stable-Diffusion-1.5	76.26(4)	75.79(4)	77.32(4)	25.76(4)	25.58(4)	25.76(3)	83.23(4)	82.64(4)	83.88(4)

Table 1: Results on MSCOCO using different metrics and different random seeds. The value under metric name is the corresponding random seed used. The value in the brackets is the ranking under the same seed.

et al., 2014), selecting 1,000 prompts from its validation split to assess general text-to-image generation capabilities following common practice (Esser et al., 2024). Additional implementation details are provided in Appendix B.

4 Experiment Results and Analysis

4.1 Analysis of Robustness

Robustness to Randomness Our investigation first examines evaluation robustness under random seed variations. Table 1 reveals that both CLIPScore and DSGScore produce inconsistent model rankings across different random seeds, demonstrating their failure to maintain robust evaluation outcomes.

Notably, this inconsistency cannot be solely attributed to narrow score margins between models. For instance, CLIPScore with seed 3407 shows a 0.10 performance gap between Pixart and SD-XL, compared to a larger 0.13 gap between Pixart and SD-1.5. Despite this greater margin, CLIPScore inconsistently ranks Pixart versus SD-1.5 while maintaining stable rankings for SD-XL.

We emphasize that our analysis does not presuppose the rankings produced by VQAScore are inherently “correct." Rather, our core argument establishes that any metric failing to maintain consistent rankings under random seed variations cannot be considered trustworthy, as there exists no reliable basis to determine which ranking might be "correct" when results fluctuate.

Furthermore, while 1000 prompts constitute a large-scale evaluation in text-to-image research, our findings reveal significant robustness failures even at this scale. This persistent inconsistency underscores fundamental challenges in current evaluation practices.

Takeaway 1: An evaluation failing to provide robust ranking under randomness should be viewed less trustworthy, like CLIPScore and DSGScore.

Robustness to Image Perturbation For our image perturbation analysis, we utilize SD-3 (selected for its superior generation performance). The evaluation employs a fixed random seed of 42, with complete results presented in Table 2.

	Avg. ΔJ_i	Max. ΔJ_i
VQAScore	0.44	10.36
CLIPScore	0.74	7.30
DSGScore	3.09	50.00

Table 2: The average and maximum performance gap between original image and perturbated image of different metrics.

The results reveal that even a minimal perturbation of 1 pixel value creates significant performance variations. While VQAScore demonstrates relatively better robustness with an average performance gap of 0.44, both CLIPScore and DSGScore exhibit unacceptably large variations, indicating fundamental robustness limitations.

More concerning are the worst-case scenarios, where this simple perturbation produces dramatic performance gaps across all metrics. Figure 1 illustrates one such failure case for VQAScore. These extreme cases are particularly problematic as they not only occur unpredictably, but also reveal potential for metric exploitation. Further, this worse case may confuse model development when visually identical inputs produce substantially different evaluations.

Takeaway 2: All three metrics fail to maintain a worst case robustness under a simple slight perturbation of image. CLIPScore and DSGScore even fail on average case, revealing a worrying fact of the trustworthiness of their evaluation result.

4.2 Analysis of Significance

We present the p-value of paired T-test and dominance ratio R to explore the significance of the comparison using different metrics in Figure 2.

We first examine the statistical significance of model comparisons. Following conventional stan-

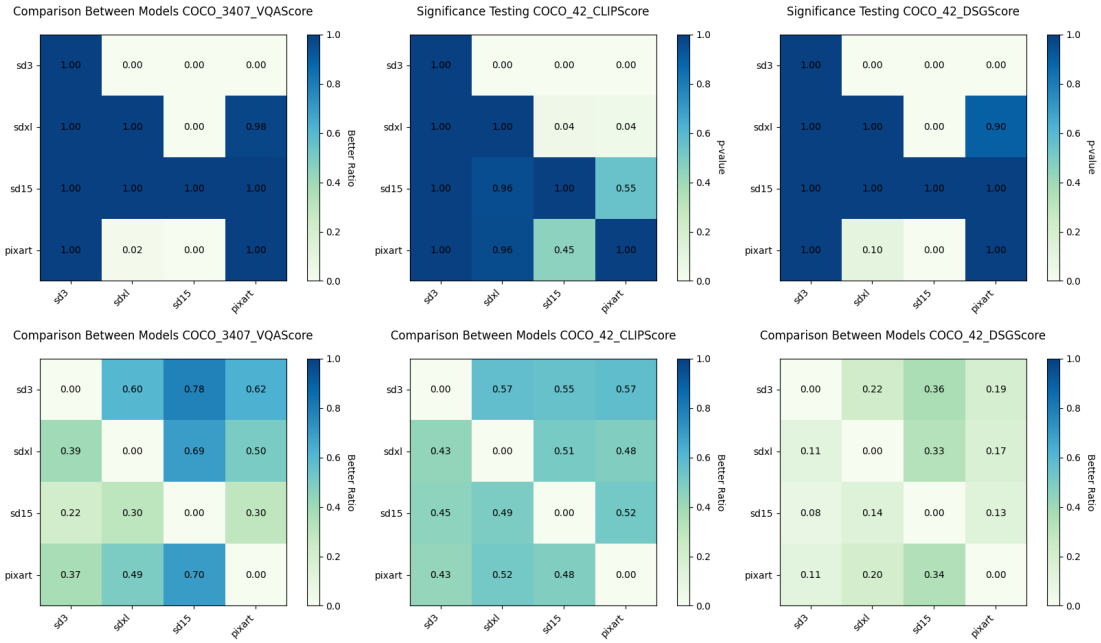

Figure 2: The T-test p-value and better ratio R between models. The value at i-th row and j-th column in a matrix represents the result of model i compared against model j. The random seed used is shown in the title of the corresponding heatmap.

dards, we consider results with p-value < 0.05 as statistically significant. Our analysis using VQAScore reveals several notable findings: SD-3 demonstrates statistically significant superiority over SD-XL, Pixart, and SD-1.5. More surprisingly, Pixart shows significant improvement over SD-XL (87.16 vs 86.01) in Table 1, seed 3407), indicating that even a score difference of 1 can reflect meaningful performance gaps between models.

This phenomenon is not unique to VQAScore. CLIPScore exhibits a similar pattern, showing SD-XL significantly superior to Pixart (25.93 vs 25.78). Similarly, DSGScore suggests potential superiority of Pixart over SD-XL ($p = 0.10$), despite their small score difference (90.52 vs 89.68). These consistent findings across metrics challenge conventional assumptions, suggesting that statistically significant improvements may occur with smaller metric differences than previously believed. So are current standards for determining meaningful model improvements excessively stringent?

To further investigate this phenomenon, we examine the dominance ratio - the probability that model i generates superior results to model j for a given prompt. We reach another shocking observation that, even if model i is significantly better than model j, the generation result of model i may not bear a large probability of being better than model j. Considering the significance derived using VQAScore as mentioned before, actually SD-XL bears a 50% probability generating a “better" result, while Pixart bears 49%, even less than SD-XL! Further, SD-3 just bears 60% probability of generating better results, while SD-XL still bears 40% probability of generating better results, even with a VQAScore gap of 4.5.

In this context, the problem is not about metrics only. The problem is, how we view “significance". If we simply view “significance" as a statistical test, we can happily accept a small improvement in metrics since it is enough to demonstrate this significance. However, if we would like to guarantee a better generation performance, it is still essential to seek an even larger metric improvement.

Takeaway 3: A small difference between metrics is enough to reveal a “significance" in statistical analysis. However, this “significance" may not be directly interpreted to actual model performance, while a trustworthy evaluation should take both aspects into account.

5 Conclusion

In this work, we introduce two important aspects, Robustness and Significance, that evaluation frameworks should focus. We conduct a wide range of experiments and reveal some important conclusions that should be taken into consideration when conducting future evaluation on image-text alignment.

Limitations

The main limitation of this work is that it does not provide a better evaluation framework to address the problems discussed in this paper. A better evaluation framework is left for future works.

References

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024a. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv*.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024b. Unified hallucination detection for multimodal large language models. *arXiv preprint arXiv:2402.03190*.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. 2023. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, volume 36, pages 52132–52152. Curran Associates, Inc.

Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. 2023. Benchmarking spatial relationships in text-to-image generation. *Preprint*, arXiv:2212.10015.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20406–20417.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Preprint*, arXiv:2307.06350.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. 2024. Evaluating and improving compositional text-to-visual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5290–5301.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*.

Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. 2023. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14277–14286.

Maitreya Patel, Tejas Gokhale, Chitta Baral, and Yezhou Yang. 2024. Conceptbed: Evaluating concept learning abilities of text-to-image diffusion models. *Preprint*, arXiv:2306.04695.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajić, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Pinelopi Papalampidi, Ira Ktena, Chris Knutsen, Cyrus Rashtchian, Anant Nawalgaria, Jordi Pont-Tuset, and Aida Nematzadeh. 2025. Revisiting text-to-image evaluation with gecko: On metrics, prompts, and human ratings. *Preprint*, arXiv:2404.16820.

Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. 2024. Conceptmix: A compositional image generation benchmark with controllable difficulty. *Preprint*, arXiv:2408.14339.

Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. 2024. What you see is what you read? improving text-image alignment evaluation. *Advances in Neural Information Processing Systems*, 36.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Related Works

There are many metrics and benchmarks focusing on evaluating image-text alignment in text-to-image generation.

CLIP-Score (Radford et al., 2021; Hessel et al., 2021) evaluates image-text alignment by computing the cosine similarity of CLIP embeddings. VQAScore (Lin et al., 2024) queries a VQA model to determine if the image corresponds to the prompt, using the "Yes" logit as the metric. T2I-CompBench (Huang et al., 2023) leverages MiniGPT-4 (Zhu et al., 2023) CoT to generate an alignment score.

Decomposition-based metrics break down text prompts into smaller components and assess the accuracy of each part. TIFA (Hu et al., 2023) generates visual questions and uses a VQA model to verify the correctness of each component. T2I-CompBench (Huang et al., 2023) employs Blip-VQA (Li et al., 2022), while DavidSceneGraph (Cho et al., 2023) and VQ^2 (Yarom et al., 2024) are similar to TIFA. MHalu-Bench (Chen et al., 2024b) builds a pipeline of tools to check the correctness of each component directly. ConceptMix(Wu et al., 2024) is a more complicated benchmark.

Some other benchmarks includes (Gokhale et al., 2023; Patel et al., 2024), GenEval (Ghosh et al., 2023).

However, all these works consider only alignment with human annotation to evaluate the validity of their evaluation. (Wiles et al., 2025) takes a step forward and points out different human evaluation template influences evaluation results, but it still focuses only on human annotation. To the best of our knowledge, we are the first to explore image-text alignment evaluation from the inner properties of the evaluation.

B Experiment Details

We use the default inference hyper-parameter of each model used. We list the details as follows:

Model Name	T	w
Stable-Diffusion-3	28	7.0
Stable-Diffusion-XL	50	5.0
Stable-Diffusion-1.5	50	7.5
PixArt-Σ-XL	20	4.5

Table 3: Details of our inference hyper-parameter. T represents total denoising steps and w represents guidance scale.