EXPLORING GPT-4 VISION FOR TEXT-TO-IMAGE SYNTHESIS EVALUATION

Xiao Cui, Qi Sun, Wengang Zhou & Houqiang Li *
University of Science and Technology of China
{cuixiao2001,qisun}@mail.ustc.edu.cn,{zhwg,lihq}@ustc.edu.cn

ABSTRACT

This paper addresses the critical need for more accurate evaluation methods in text-to-image synthesis. While the standard CLIPScore metric can reflect text-image alignment to some extent, it often falls short in consistency with human perception. We propose the use of GPT-4 Vision as a novel evaluative standard, capable of interpreting text and image nuances akin to human cognition. Our study focuses on the pivotal role of prompt design in maximizing GPT-4 Vision's effectiveness, presenting a systematic discussion for prompt construction. Empirical evaluations demonstrate that GPT-4 Vision, augmented by our prompt-design strategy, aligns more closely with human judgment.

1 INTRODUCTION

Recent developments in text-to-image synthesis have been significant, with various models achieving remarkable results (Ramesh et al., 2021; 2022; Ding et al., 2021; Rombach et al., 2022; Jahn et al., 2021; Nichol et al., 2021). The predominant method for evaluating text-image alignment in these models, primarily the CLIPScore metric (Li et al., 2019; Yin et al., 2019; Lee et al., 2023), has demonstrated notable inconsistencies in aligning with human perception (Otani et al., 2023; Xu et al., 2023). This inconsistency underscores a need for a more effective evaluation approach.

Addressing this gap, our research introduces GPT-4 Vision (OpenAI, 2023) as an innovative alternative. Unlike conventional metrics, GPT-4 Vision leverages advanced algorithms to interpret textual and visual cues more akin to human cognition, thus offering a potentially more accurate assessment.

The efficacy of GPT-4 Vision, however, is intricately tied to the design of the input prompts. Our study delves into this aspect, presenting a systematic methodology for crafting prompts that maximize GPT-4 Vision's evaluative capabilities. This approach not only enhances the model's performance accuracy but also provides foundational insights for future explorations in the field.

Our empirical findings indicate that GPT-4 Vision, with its refined prompt-design strategy, serves as a more aligned and autonomous assessment tool with human evaluative standards, marking a step forward in text-to-image synthesis evaluation.

2 METHOD AND EXPERIMENTS

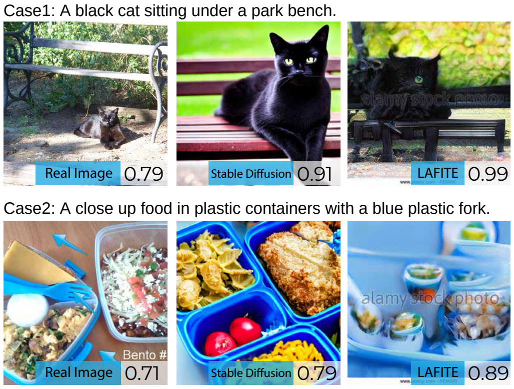
Method For a given input text, the model under evaluation generates an image. This image, alongside a formulated text prompt derived from a structured template, is submitted to GPT-4 Vision (OpenAI, 2023) for scoring. Our experimentation includes four distinct types of prompts: a basic version with general descriptors, a label-enhanced version incorporating more precise and specific categorical labels, a question-enhanced version focusing on selected key aspects, and a comprehensive version thoroughly detailing all relevant aspects for GPT-4 Vision's evaluation. The specifics of these prompt templates are comprehensively detailed in the Appendix, providing clarity on the evaluation process and allowing for reproducibility of our methods.

[*] Xiao Cui and Qi Sun contributed equally to the study.

Methods	**SD**		**GLIDE**		**Relative**
	Pearson	Spearman	Pearson	Spearman	Matthews
CLIP	0.24	0.23	0.20	0.24	0.43
Ours (basic)	0.39	0.33	0.58	0.54	0.63
Ours (label-enhanced)	0.42	0.38	0.62	0.62	0.61
Ours (question-enhanced)	0.42	0.42	0.66	0.62	0.68
Ours (comprehensive)	**0.49**	**0.48**	**0.69**	**0.65**	**0.70**

Table 1: Quantitative analysis of GPT-4 Vision with various prompt types versus CLIP for evaluation consistency with human assessment. We report correlations between the scores produced by each model and human evaluations, along with a relative scoring strategy.

Settings Our study aimed to assess the alignment between various evaluators and human judgment. We employed 50 text labels from the COCO-caption dataset (Lin et al., 2014; Chen et al., 2015) to generate images using Stable Diffusion (Rombach et al., 2022) and GLIDE (Nichol et al., 2021). Thirty participants, with varied backgrounds, were involved in the human evaluation study. They responded to 100 questions that were consistent with those used in the automated model evaluation, ensuring a standardized comparison. The evaluation comprised two parts. First, we employed both the Pearson correlation coefficient (Cohen et al., 2009) and Spearman's rank correlation coefficient (Sedgwick, 2014; Zar, 2005) to calculate the correlation between each automated evaluator and human evaluation. Furthermore, a relative scoring system was implemented: an image generated by Stable Diffusion receiving a higher score than GLIDE was assigned a 1, and a 0 otherwise. This methodology was paralleled in the human evaluation segment. The alignment between automated evaluators' and human judgments was quantified using the Matthews correlation coefficient (Chicco et al., 2021; Yao & Shepperd, 2020).


Case1: A black cat sitting under a park bench.

Case2: A close up food in plastic containers with a blue plastic fork.

Figure 1: Failure cases on CLIPScore.

Prompt	Case	L	M	R
Basic	1	4	5	5
	2	5	4	3
Label-enhanced	1	5	4	5
	2	5	5	2
Question-enhanced	1	5	4	4
	2	5	5	2
Comprehensive	1	5	2	2
	2	5	3	2

Table 2: Our evaluation scores. For each subfigure in Fig. 1, our evaluators assigned scores using different prompts. Notations: L - Left; M - Middle; R - Right.

Results As illustrated in Table 1, a higher CLIPScore does not necessarily equate to better alignment. Our experiments demonstrate that GPT-4 Vision significantly outperforms CLIP across all evaluated correlation matrices. This performance enhancement is particularly evident in scenarios involving specifically designed input prompts. Figure 1 presents some failure cases in CLIPScore evaluations. Further, as Table 2 demonstrates, our methodology, which incorporates a comprehensive design of prompts, not only identifies discrepancies between text and image with high accuracy but also aligns closely with human perceptual judgments. More results are shown in the Appendix.

Discussion An integral aspect of employing GPT-4 Vision (OpenAI, 2023) for text-to-image synthesis evaluation is the design of the prompts. It is crucial that these prompts possess explicit clarity to minimize ambiguity and ensure reproducibility. Vague or overly broad prompts can lead to significant randomness in the evaluation process, undermining the standard's reliability. Additionally, the prompts must be crafted with precision to direct GPT-4 Vision's attention to specific, relevant aspects of the image. This precision is vital to prevent the overlooking of crucial details, ensuring a comprehensive and accurate assessment of the text-to-image alignment. With suitable prompts, GPT-4 Version can serve as a new standard for text-to-image synthesis evaluation which has better consistency with human perception.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

References

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Davide Chicco, Niklas Tötsch, and Giuseppe Jurman. The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining*, 2021.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, 2009.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. CogView: Mastering text-to-image generation via transformers. In *NeurIPS*, 2021.

Manuel Jahn, Robin Rombach, and Björn Ommer. High-resolution complex scene synthesis with transformers. In *CVPR*, 2021.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *arXiv preprint arXiv:2311.04287*, 2023.

Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *CVPR*, 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *CVPR*, 2023.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022.

Philip Sedgwick. Spearman's rank correlation coefficient. *British Medical Journal*, 2014.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022.

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. *arXiv preprint arXiv:2304.05977*, 2023.

Jingxiu Yao and Martin Shepperd. Assessing software defection prediction performance: Why using the matthews correlation coefficient matters. In *Proceedings of the 24th International Conference on Evaluation and Assessment in Software Engineering*, 2020.

Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics disentangling for text-to-image generation. In *CVPR*, 2019.

Jerrold H Zar. Spearman rank correlation. *Encyclopedia of biostatistics*, 2005.

Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. GPT-4V (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*, 2023.

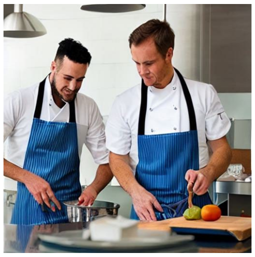
A Appendix

A.1 Prompt

Here we present a detailed exposition of our prompt template architecture, which encompasses four distinct designs: a basic template, a label-enhanced template, a question-enhanced template, and a comprehensive template.

Basic design How well does the image match the description? 1.Does not match at all 2. Matches just slightly 3. Matches somehow 4. Matches almost exactly 5. Matches exactly

Label-enhanced version How well does the image match the description? 1. Does not match at all 2. Has signifcant discrepancies 3. Has several minor discrepancies 4. Has a few minor discrepancies 5. Matches exactly



Input: How well does the image match the description "Two men wearing aprons working in a commercial-style kitchen."? 1. Does not match at all 2. Has signifcant discrepancies 3. Has several minor discrepancies 4. Has a few minor discrepancies 5. Matches exactly. Only output the answer.

GPT-4 Vision Output: 5. Matches exactly.

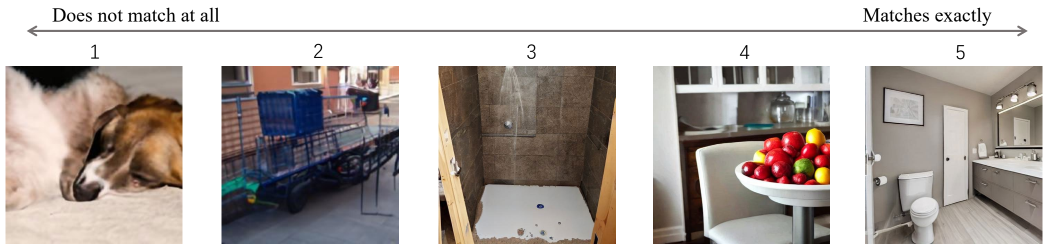
Figure 2: An example of GPT-4 Vision for evaluating text-image alignment.

Question-enhance version How well does the image match the description? Carefully examine location and each element. 1.Does not match at all 2. Matches just slightly 3. Matches somehow 4. Matches almost exactly 5. Matches exactly

Comprehensive design Rate how well the image matches the following description, considering the presence, position, and characteristics of each described element. Examine the image for [Element 1], [Element 2], ..., and [Element N], noting their [color/shape/size/number/position]. Assess the [background/foreground/context] for any additional elements or discrepancies. Provide a score based on the following scale: 1.The image does not contain any of the described elements. 2.The image contains the elements, but there are significant differences in major characteristics such as color, size, or number. 3.The image includes the elements with some variations in less crucial characteristics like exact position or orientation. 4.The image shows the elements with only minor variations that do not significantly alter the overall impression. 5.The image matches the description exactly with all elements present and accurately depicted as described.

Figure 2 illustrates a representative example, showcasing an input image and corresponding input text, alongside the resultant output from the GPT-4 Vision model.

A.2 Our Automatic ratings on cases



A baby is laying down with a teddy bear.

A person with a shopping cart on a city street

A bathroom with a walk in shower currently under repair.

The dining table near the kitchen has a bowl of fruit on it.

A bathroom with a white toilet sitting next to a bathroom sink.

Figure 3: Generated images and their automatic ratings (using GPT-4 Vision with comprehensive prompt design) of text-image alignment.

A.3 Failure prompt

Trial	Aspect	Left	Middle	Right
1	Relevance	37	38	36
	Object Accuracy	28	27	29
2	Relevance	38	35	40
	Object Accuracy	28	27	28
3	Relevance	38	40	40
	Object Accuracy	27	30	27

Table 3: Evaluation scores using failure prompt. Each sub-figure in Fig. 1's second case was subjected to three experimental trails using faulty prompts. The results indicate considerable variability and a lack of precision.

Here we evaluate the prompt methodology employed in (Zhang et al., 2023): 'Carefully assess the generated image in terms of relevance to the prompt and object accuracy. Use the following criteria to guide your evaluation: with Relevance (0-40 points), Object Accuracy (0-30 points).' Table 3 demonstrates that the outputs from GPT-4 Vision display substantial variability and a marked misalignment with human perception. These limitations are primarily attributed to the lack of specificity and clarity in the scoring criteria provided by the prompt, contributing to the observed inconsistency in the results. Importantly, in the methodology presented in our main text, the evaluator's scoring results do not exhibit such randomness, demonstrating the robustness and reliability of our approach.

A.4 Details on text-to-image models

In align with the previous work (Otani et al., 2023), we conduct GPT4-V evaluations on 2 different text-to-image models: Stable Diffusion (Rombach et al., 2022) and GLIDE (Nichol et al., 2021).

GLIDE is a diffusion-based text-to-image model with classifier-free guidance, and we adopt the released official notebook to produce our samples[1].

[1] https://github.com/openai/glide-text2im

Stable Diffusion V1-5 is another diffusion-based text-to-image model based on latent diffusion, trained on large-scale text-image dataset LAION (Schuhmann et al., 2022), and we generate the images with the popular huggingface diffusers (von Platen et al., 2022)[2].

All samples necessary to reproduce the evaluation are released on the provided anonymous link [3].

[2] https://huggingface.co/runwayml/stable-diffusion-v1-5

[3] https://anonymous.4open.science/r/GPT4-V-text2img-evaluation-EF70