Image Regeneration: Evaluating Text-to-Image Model via Generating Identical Image with Multimodal Large Language Models

Chutian Meng, Fan Ma, Jiaxu Miao, Chi Zhang, Yi Yang, Yueting Zhuang*

College of Computer Science and Technology, Zhejiang University
{12321215, mafan, jiaxumiao, 12321216, yangyics, yzhuang}@zju.edu.cn

Abstract

Diffusion models have revitalized the image generation domain and play crucial roles in both academic research and artistic expression. With the emergence of new diffusion models, the evaluation of the performance of text-to-image models has become increasingly important. Current metrics focus on directly matching the input text with the generated image, but due to cross-modal information asymmetry, this leads to unreliable or incomplete assessment results. Motivated by this, we introduce the Image Regeneration task in this study to assess text-to-image models by tasking the T2I model with generating an image according to the reference image. We use GPT4V to bridge the gap between the reference image and the text input for the T2I model, allowing T2I models to understand the content of the image. This evaluation process is simplified, as comparisons between the generated image and the reference image are straightforward. Two regeneration datasets spanning content-diverse and style-diverse evaluation dataset are introduced to evaluate the leading diffusion models currently available. Additionally, we present ImageRepainter framework to enhance the quality of generated images by improving content comprehension via MLLM guided iterative generation and revision. Our comprehensive experiments have showcased the effectiveness of this framework in assessing the generative capabilities of models. By leveraging MLLM, we have demonstrated that a robust T2M can produce images more closely resembling the reference image.

Code&Datasets: https://github.com/SPEEDSRER/Image-Regeneration

Introduction

With the rise of generative AI, there has been a surge in the development of text-to-image (Rombach et al. 2022a; Podell et al. 2023; Ramesh et al. 2022; Saharia et al. 2022; Yu et al. 2022) and text-to-video (Khachatryan et al. 2023; Liu et al. 2024) models in recent years. The primary goal of such algorithms is to generate visual content based on user prompts. These technologies have broad applications (Yang et al. 2024b), such as style transfer (Zhang et al. 2023; Wang, Zhao, and Xing 2023), image editing (Gal et al. 2022; Ruiz et al. 2023a,b; Zhou et al. 2024b,a; Liang et al. 2024), and more, holding significant importance in academic and creative domains.

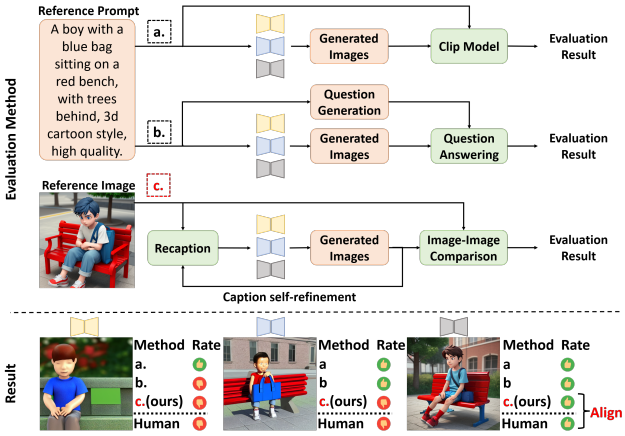


Figure 1: Workflow comparison among (a) pre-trained model evaluation, (b) QG&QA(Question Generation & Question Answering) evaluation, (c) image regeneration task evaluation, where our approach (c) achieves better alignment with human cognition.

Despite rapid advancements in related algorithms and applications, a research gap remains in evaluating the quality of these generative models. Current T2I model evaluation focuses on two modalities: text input and image output. For example, CLIP score (Radford et al. 2021) is widely used to measure the semantic information within images based on user prompts. However, the CLIP score has lower sensitivity to fine-grained details and visual variances and cannot reflect the quality of generated images. Similarly, the recently popular evaluation method QG&QA (Question Generation and Question Answering) generates several questions and answers related to the generated image using the textual input and employs VQA methods to compare and evaluate the image, such as T2I-CompBench (Huang et al. 2023). Although such methods have achieved some results in measuring the consistency between text and image content, they cannot effectively assess the model's overall performance under complex prompt conditions.

We can derive two insights from the current evalua-

*Corresponding author.

tion methods. First, current methods focus on two different modalities, text input and image output, which inherently involve differences and information asymmetry between modalities, making evaluation challenging. Second, a good generative model should perform well in complex, real-world scenarios. Current methods focus on evaluating single attributes, but they lack comprehensive assessment for complex real-world scenarios involving multiple conditions.

Motivated by these insights, we propose **Image Regeneration** task, similar to "Painting Reproduction", given a reference image, we require the model under evaluation to generate an image based on it. The generated image is then assessed by comparing it to the reference image to determine the model's generation ability. Figure 1 compares the workflow of image regeneration task and current metrics. The reference image is more informative than text prompts and aligns the output modalities, leading to more reasonable evaluation results. Since the generative model uses text inputs, we have developed an MLLM(Yang et al. 2023a; Nyberg et al. 2021) based method to convert from image to text input. We propose a framework ImageRepainter for evaluating the quality of text-to-image models based on image regeneration task. The framework involves two stages. Specifically: **(1) Image understanding:** Firstly, based on MLLMs, the image information is organized to generate a tree-like structure called the image understanding tree (IUT), and then text prompts are generated using the information from it. **(2) Iterative generation:** Iterative exploration (Yang et al. 2023b, 2024a; Wang et al. 2023) is typically involved in T2I generation for better images. This stage includes 4 parts: prompt generation/revision, image generation, image selection, and feedback generation. Furthermore, We introduce two benchmarks respectively designed for the evaluation of the content and style of the generated results.

The contributions if this work can be summarized as:

- **Novel T2I Model Evaluation Task:** This framework's conception is rooted in the concept of "Painting Reproduction" which naturally aligns with the human judgment methodology. The task is able to measure the overall **generative capability** of the T2I model as well as reflect its **generative speciality**.
- **Effective Image Comprehensive Mechanism:** We introduce Image Understanding Tree(IUT) to enhance MLLMs' mutimodel ability. By defining fundamental rules, we incentivize MLLMs to interact with images and summarize the image information in a hierarchical tree structure, namely IUT.
- **Two Diverse Benchmarks:** We propose two benchmarks respectively designed for the evaluation of the **content** and **style** of the generated results. The experimental results indicate that our benchmark aligns with human perception in assessing the generation capabilities of T2I models.

Related Work

Evaluation of Generation Tasks

Existing metrics for text-to-image generation can be categorized into fidelity assessment, alignment assessment, and LLM-based metrics (Huang et al. 2023). Traditional metrics such as Inception Score (IS) (Salimans et al. 2016) and Frechet Inception Distance (FID) (Heusel et al. 2018) are commonly used to evaluate the fidelity of synthesized images. For evaluating image-text alignment, CLIP (Radford et al. 2021) and BLIP2 (Li et al. 2023) are typically used for semantic matching between text and image. Recently, more and more work leverages the powerful reasoning capabilities of LLMs (Yang et al. 2023b, 2024a) for evaluation. However, the measurement of text-to-image generation does not intuitively reflect human comparisons and perceptions of visual information. Therefore, we propose an evaluation framework for the text-to-image model based on the image regeneration task, simulating the form of human painting reproduction, which provides a natural and reasonable judgment for T2I model quality.

LLM-driven Automatic System

In the field of natural language processing (NLP), there has been a significant transformation with the emergence of LLMs (Chowdhery et al. 2022; Ouyang et al. 2022; Touvron et al. 2023), which have demonstrated significant capabilities in interacting with humans through conversational interfaces. The "Chain-of-Thought" (CoT) framework (Kojima et al. 2023; Wei et al. 2023; Zhang et al. 2022) opens the door to further enhance the capabilities of LLMs, which guides LLMs to progressively generate answers, aiming to obtain higher-quality response. Recent research has pioneered novel methodologies by integrating external tools or models with Large Language Models (LLMs). For example, Toolformer (Schick et al. 2023) facilitates LLMs' access to external tools via API tags. Visual ChatGPT (Wu et al. 2023) and HuggingGPT (Shen et al. 2023) have broadened the scope of LLMs by enabling them to leverage other models for tasks extending beyond linguistic domains. Furthermore, PromptBreeder (Fernando et al. 2023) and Idea2Img (Yang et al. 2023b) frameworks respectively utilize LLMs for automated optimization of text prompt and image design. Inspired by these efforts, we embrace the concept of LLMs as multifunctional tools and utilize this paradigm to construct an iterative framework for evaluating the quality of generative models with the image regeneration task.

Methodology

We introduces ImageRepainter, shown in Figure 2, a T2I model evaluation framework based on the image regeneration task. It employs MLLMs facilitating T2I models to generate images based on a specific reference image in terms of content, style, and other aspects. Finally, the framework utilizes the image-to-image metrics between the generated images and the reference images as the standard for evaluating the quality of the generation model. Similar to the concept of human "Painting Reproduction", it is easier for humans to make comparison and judgement within the same modality.

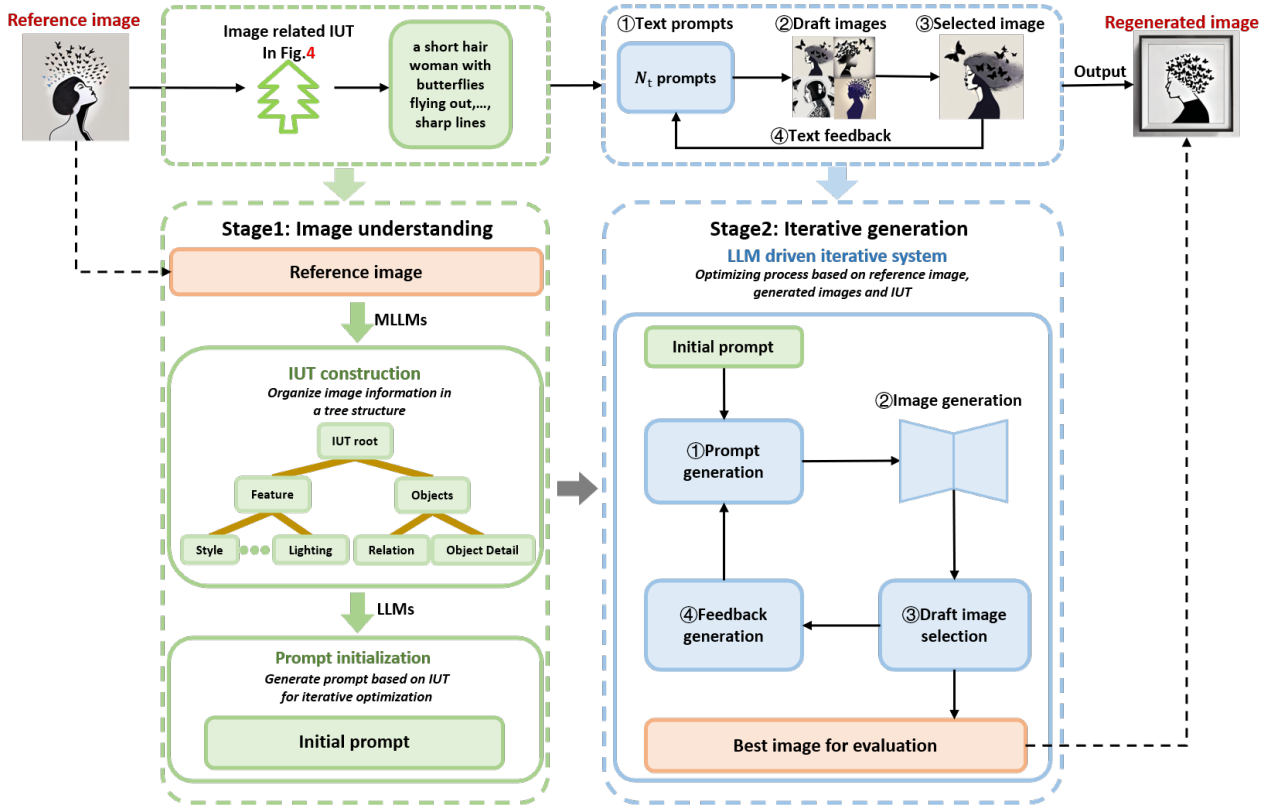


Figure 2: **Overview of ImageRepainter**. The framework consists of two stages: **image understanding** and **iterative generation**. These stages are displayed from left to right and interact continuously with LLM.

Image Understanding

Image understanding is the first stage of this framework aiming at generating a high quality description of the reference image. The CLIP-interrogator model (Li et al. 2022; Radford et al. 2021) in can generate stable diffusion prompts associated with the image input, thus producing image understanding. We directly employ the CLIP-interrogator for image regeneration tasks, as shown in Figure 3. The prompt produced may contain jumbled text and lacks accuracy. Therefore, the CLIP-interrogator cannot serve as a ideal method for image understanding.

In this regard, we utilize MLLMs for image understanding, since they exhibit strong capabilities and align well with human cognition. We introduce Image Understanding Tree (IUT), in order to organize the information of an image in a tree structure, as this prevents redundancy and allows for a clear delineation of features at various levels of granularity. Constructing the IUT requires the use of multimodal large language model M(GPT4v) to analyze the reference image. We design templates to generate JSON format output for its standardization.

As shown in Figure 4, given a portrait image I_{ref}, the first step is to generate a caption p_{cap} of the image as a base prompt. Then, we guide the MLLM to extract the overall features of the image $f_{I_{ref}}$, the objects within the image $\{x_0, ..., x_{k-1}\}$, and the relationships between the objects R_{obj} from the image using text template T_{ext}

$$\{f_{I_{ref}}, x_0, ..., x_{k-1}, R_{obj}\} = M(I_{ref}, T_{ext}). \quad (1)$$

Subsequently, more detailed information is extracted for each object in the image. After using LLM for automated questioning with text template T_{obj}, detailed information about the respective objects is obtained.

$$\{f^0_{x_i}, ..., f^{n-1}_{x_i}\} = M(I_{ref}, x_i, T_{obj}) \quad (2)$$

, where $f^j_{x_i}$ represents the jth feature of the object x_i. Fi-

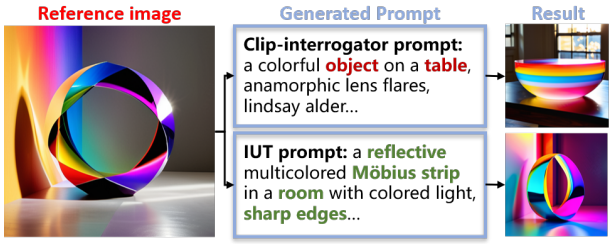


Figure 3: Examples of the generated images by using the prompt from CLIP-interrogator and our proposed IUT. We can observe that the accuracy of the information described in the prompts generated by the CLIP-interrogator is insufficient, leading to unsatisfactory results due to incomplete information.

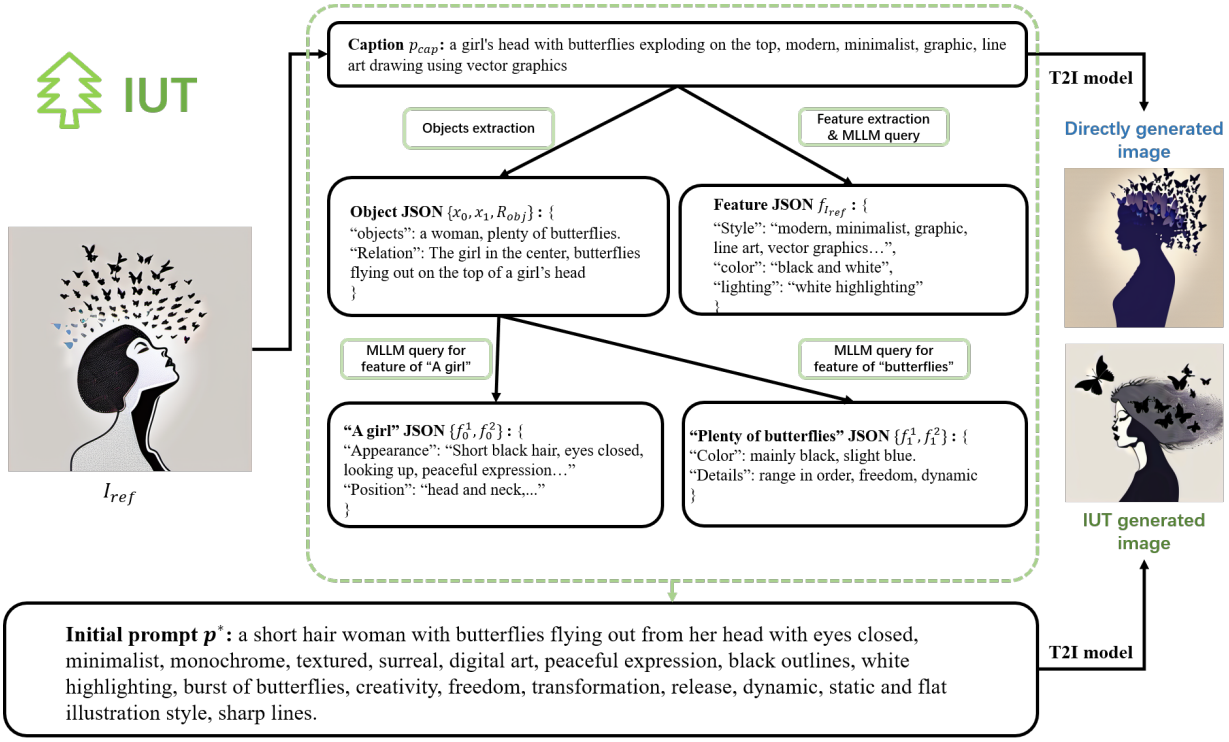


Figure 4: An example IUT construction, which shows that IUT capture more information such as color and facial details of the image than the direct caption.

nally, the initial prompt p^* is generated based on the IUT of the reference image using LLM.

Iterative Generation

After generating prompt p^* from the image understanding module, for the initial iterations, the prompt generation module is responsible for expanding the prompt p^* into N_1 synonymous prompts to participate in the iterations, where N_t represents the number of the prompts in the t^{th} iteration. The reason for generating multiple synonymous prompts is due to the T2I model's bias in understanding synonymous words, which can lead to situations where a good prompt generates a bad image.

The iterative process comprises 4 parts: image generation, image selection, feedback generation, and prompt revision.

Image generation. In the t^{th} iteration, the framework input the N_t prompts obtained from the prompt generation part into the T2I model under evaluation, resulting in N_t corresponding image outputs.

Image selection. CLIP, DINOv2, GPT4v is used to assess the similarity between N_t images and the reference image, selecting the highest scoring image and prompt. If the current iteration t reaches the maximum iteration T, the images from this iteration are considered as the model's regeneration result of the reference image; otherwise, the iteration continues. This part is designed to retain relatively stable performance and generate high-quality images. CLIP and DINOv2 metrics are capable of measuring coarse-grained semantic and visual information in images, while GPT4v captures fine-grained content and perceputual information.

Feedback generation. Feedback generation is to provide guidance for modifying the prompt. Text feedback F is generated based on the differences between the current image and the reference image, as well as the previously constructed IUT.

Prompt revision. For intermediate iteration t, prompt modification is carried out based on the best-performing prompt after evaluation in the current iteration and the text feedback F, generating N_{t+1} prompts for image generation. In each iteration, we only modify one aspect of the prompt, since we observe in practice that it can make the LLM more "focused" and generate better prompt results.

Experiments

Experiments Setting

Task Definition: In the **image regeneration** task, T2I models are required to generate an image according to a reference image, akin to the process of human image repainting. It provides a more intuitive assessment compared to directly evaluating the alignment between the T2I model's input text and output image.

Human evaluation: When evaluating T2I models, the quality of the judgment method is determined by its alignment with human assessments. Our user study references the Likert scale human evaluation template ImagenHub (Ku et al.

Model	ImageRepainter				User study		T2I-comp	HPSv2/Pickscore
	CLIP(%)	DINO(%)	GPT4-con	GPT4-per	consistency	perceptual		
SD1.4	89.98	88.37	0.5500	0.4460	0.4216	0.3682	0.3080	0.2563/0.0871
SD1.5	90.17	90.33	0.5660	0.4760	0.4338	0.3976	0.3315	0.2584/0.1015
SD1.5-DPO	91.08	93.57	0.5560	0.6840	0.4544	0.6388	0.3392	0.2614/0.1390
SD2.0	90.79	91.68	0.6060	0.5860	0.4960	0.4626	0.3386	0.2577/0.0787
SD2.0-Inpaint	90.67	92.32	0.6220	0.6020	0.4892	0.4920	0.3560	0.2611/0.0854
SDXL1.0	90.17	92.41	0.7600	0.6600	0.6726	0.6448	**0.4091**	0.2565/0.0999
Juggernautv1	93.37	95.27	0.7120	0.7740	0.6472	0.8072	0.3476	0.2705/**0.2118**
Juggernautv9	**93.79**	**95.34**	**0.7700**	**0.8820**	**0.7056**	**0.8590**	0.3764	**0.2731**/0.1967

Table 1: The evaluation result of our proposed ImageRepainter framework, T2I-CompBench, and user study. GPT-con and GPT-per represents the content consistency and perceptual quality evaluated by GPT4v.

2024). Specifically, we ask annotators to rate both the content consistency between the generated image and the text prompt, and the perceptual quality of the image, on a scale from 1 to 5. For each model, we randomly select 50 text-image pairs for evaluation, with each pair being rated by 5 human annotators. A total of 40 participants are involved, with each annotator rating 50 text-image pairs. We normalize and present the two-dimensional scores for each model in Table 1.

Baseline for T2I model evaluation: We apply T2I-CompBench (Huang et al. 2023) 3-in-1 evaluation as the baseline for content consistency evaluation which takes attribute binding, object relationship into consideration. HPSv2 score and PickScore serve as the baseline for perceptual quality evaluation.

Baseline for image understanding: We apply CLIP-interrogator as the baseline for our proposed image understanding method. The CLIP-interrogator can generate text prompts for a given image input, which in turn can generate images similar to the input. In this experiment, the CLIP-interrogator is used as a baseline to assess whether the framework is capable of completing the image regeneration task. By quantitatively comparing with the evaluation metrics of the CLIP-interrogator, we aim to demonstrate that our proposed framework can provide a more accurate and superior understanding of images.

Implementation Details: We set iteration rounds $T = 4$. We define a queue of iterative elements, where each round iterates over one element in the queue, including overall image, style, color, and detailed content in sequence.

Models: In the experiment, we utilize GPT4v (202 2023) as MLLM and employ ChatGPT to handle pure text tasks in order to save resources, specifically the text-davinci-003 version. To control the LLM's response, we utilize JSON format (Shen et al. 2023) to constrain the text output of the LLM. For the generation model used in our experiment, we employed the stable diffusion officially released model *SD1.4*, *SD1.5*, *SD2.0*, *SD2.0-inpainting* (Rombach et al. 2022b), the state-of-the-art *SDXL1.0* (Podell et al. 2023), and popular models from the open community to assess the quality of the generation model within our framework. In comparison to the officially released base models, custom models can generate relatively stable quality images. However, due to fine-tuning, the diversity in content or style generated by these models might be relatively diminished. Using multiple models can demonstrate the effectiveness of our proposed evaluation framework. *SD1.5-DPO* (Wallace et al. 2023) is a fine-tuned version of *SD1.5* by directly optimizing on human comparison data. The $JuggernautXL_{v9}$ and $JuggernautXL_{v1}$ (KandooAI 2024) are the most popular and effective models from the Civitai community.

Evaluation Datasets: For quantitative evaluation, we constructed two benchmarks, respectively designed for the evaluation of content and style of the generated results. The **style-diverse benchmark** consists of 200 text-image samples with 10 different styles. The **content-diverse benchmark** consists of 100 samples with 4 different types of content. The style-diverse benchmark is composed of 20 manually selected captions combined with descriptions of 10 style categories, and normalized and synonym-transformed using ChatGPT. The content-diverse benchmark, on the other hand, is manually collected through an open-source creation platform and normalized using ChatGPT. The data distribution of the two benchmarks is illustrated in Figure 5.

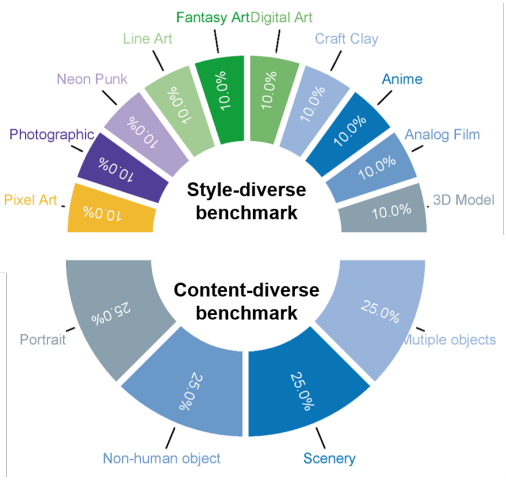


Figure 5: The distribution of style-diverse benchmark and content-diverse benchmark.

Evaluation Metrics: We use CLIP (Radford et al. 2021), DINOv2 (Oquab et al. 2024), and GPT4v. CLIP score is ca-

pable of evaluating the semantic information between images. DINOv2 metric, compared to the CLIP metric, is more sensitive to visual information such as lighting and color tones in images. GPT-4V scores are given on a scale from 1 to 5 for two dimensions: content consistency between the reference image and the generated image, and perceptual quality.

Evaluating T2I Models

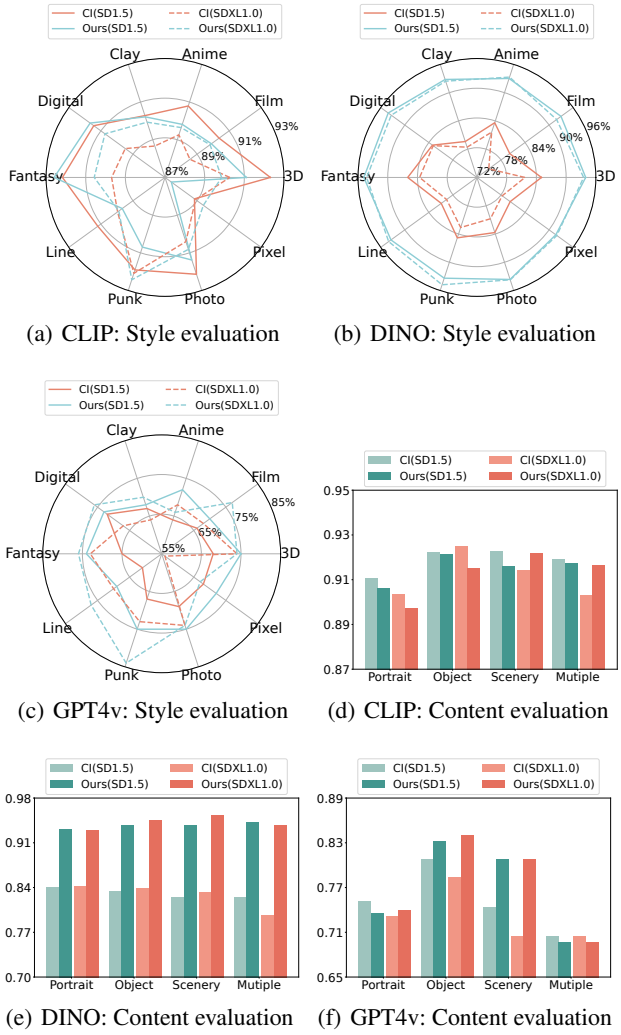
We use the content-diverse benchmark for evaluation. The results are shown in Table 1. It can be observed that ImageRepainter evaluation aligns more closely with human annotations. In terms of content consistency, the SDXL1.0 model performs best in the T2I-Compbench evaluation, though it differs from human judgments. On the perceptual level, ImageRepainter outperforms PickScore. Both our method and HPSv2 scores align with human judgments, but our method better reflects the perceptual quality differences of the models and also provides interpretable text (provided in supplementary materials).

To more intuitively demonstrate the model's comprehensive generation capabilities, we list random selected cases of ImageRepainter. As shown in Figure 6, based on visual perception, it can be observed that $JuggernautXL_{v9}$ exhibits the strongest generalization and generation capabilities, consistent with our proposed ImageRepainter evaluation results, which indicates that our proposed evaluating method better aligns with human cognition.

Figure 6: Cases intuitively demonstrate the generation capabilities of T2I models, showing that our evaluation of the four T2I models is reasonable.

Enhancement of Image Understanding

In order to prove that ImageRepainter has a correct and superior understanding of images, we performed image regeneration experiments using CLIP-interrogator as a control. We utilize content-diverse and style-diverse datasets to evaluate ImageRepainter's performance across various types of



(a) CLIP: Style evaluation (b) DINO: Style evaluation

(c) GPT4v: Style evaluation (d) CLIP: Content evaluation

(e) DINO: Content evaluation (f) GPT4v: Content evaluation

Figure 7: The ImageRepainter performs better than the CLIP-interrogator across various styles. There is a significant improvement in both the DINOv2 and GPTv4 metrics.

image inputs. We employ *SD1.5* and *SDXL1.0* as the T2I model in the framework. Figure 7 compares the CLIP, DINOv2, GPT4v metrics of the images generated by the CLIP-interrogator and ImageRepainter.

We can observe that using ImageRepainter for image regeneration tasks, the images obtained have a significant advantage over those obtained through the CLIP-interrogator for caption-based images, as judged by the criteria of DINOv2 and GPT4v. This demonstrates the effectiveness and superiority of employing the LLM-driven image regeneration and prompt iteration.

Furthermore, we can observe models' generation speciality from the results, which indicate that the *SD1.5* model and the *SDXL1.0* model perform relatively evenly across various type of styles. As for content-diverse evaluation, we observe that both the *SD1.5* and *SDXL1.0* models have limited abilities in generating portraits and multi-object object images.

Despite our framework's ability to iteratively generate better descriptions for images, the generation results remain unsatisfactory.

Ablation Study

We conduct ablation study on image regeneration task itself, IUT, and iterative process.

Effectiveness of Image Regeneration: To prove the effectiveness of our method, we conduct experiments on directly using GPT-4v for text-image matching on content-diverse dataset. As shown in **Table 2** and **Figure 8**, we can observe that directly using GPT-4v for text-image matching **lacks distinguishability** of different models compare to image regeneration task. It may stem from the fact that MLLM performs better when comparing within the same modality, while it cannot fully leverage the capabilities of large models for different modalities such as image and text. It also demonstrates the potential of the ImageRepainter evaluation method alongside the development of MLLMs.

Model	Direct GPT4v	ImageRepainter	User study
SD1.5	0.6620	0.5210	0.4157
SDXL1.0	**0.7440**	0.7100	0.6587
juggerv1	0.6960	0.7430	0.7272
juggerv9	0.7220	**0.8260**	**0.7823**

Table 2: Comparison between the ImageRepainter method and direct employment of GPT-4v for text-image evaluation.

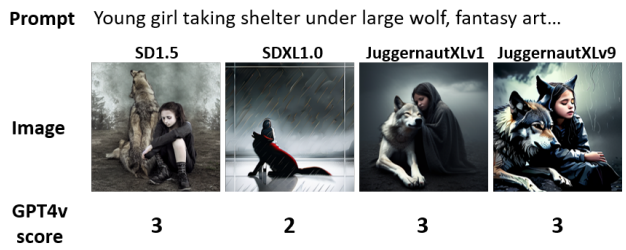


Figure 8: Direct text-image evalution via GPT4v.

Impact of IUT: In this section, we designed experiments to verify the effectiveness of the proposed IUT on image understanding and generation. Table 3 presents the results of prompt generation directly from images and prompt generation incorporating IUT. The experimental results indicate that IUT is more effective to relatively high quality models, since they exhibit strong and accurate execution of input text.

Impact of Iteration Rounds: In this section, we qualitatively demonstrate the impact of iteration rounds on the quality of image generation. Figure 9 presents several examples. We can observe that when using the higher-quality $JuggernautXL_{v9}$ model, the iterations have a less noticeable impact on the improvement of image quality. Conversely, for the relatively weaker quality *SDXL1.0* model, iterations have a more significant impact on the image quality improvement. This is because the higher-quality model generates images more consistently, while *SDXL1.0* may be influenced more by the seed and requires repeated iterations to obtain a relatively stable prompt performance.

Model	Method	CLIP(%)	DINO(%)	GPT4v(%)
$JuggerXL_{v9}$	Direct	93.66	94.83	76.8
	IUT	**95.71**(+2.05)	**95.71**(+0.88)	**84.4**(+7.6)
$SDXL1.0$	Direct	**90.18**	90.41	62.8
	IUT	90.17(−0.01)	**91.48**(+1.07)	**66.0**(+3.2)

Table 3: Quantitative results on the effectiveness of our proposed image understanding template IUT. Method "Direct" generates the initial prompt directly from the reference image using MLLM. Method "IUT" generates the initial prompt based on IUT.

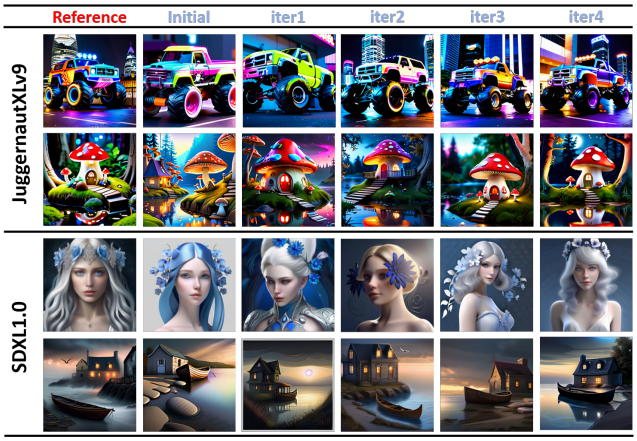


Figure 9: Ablation study of iteration rounds. The result indicates that the iteration plays a more important role in the models with relatively low quality. The models with high-quality can generate similar and good images with fewer iterations.

Conclusion

In this paper, to fill the research gap in the evaluation of generative models, we propose ImageRepainter, an LLM-driven framework for assessing the quality of text-to-image models with **image regeneration** task. The framework iteratively generate high-quality images to explore the generation capability of T2I models and evaluate the model's generation effectiveness based on a visual-to-visual intuitive understanding. The visual-to-visual assessment in this paper is better compared to current text-to-visual assessments because the former is more insensitive to fine-grained information and relatively intuitive in terms of human perception. Additionally, our ImageRepainter can contribute to the AIGC community, facilitating creative work, and can also be applied to other meaningful tasks such as dataset augmentation, demonstrating strong extensibility. For future work, we will continue to improve this new framework to incorporate more complex input conditions and further explore effective methods for assessing the quality of generative models.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (U2336212) and National Key R&D Program of China under Grant 2022ZD0160101.

References

2023. GPT-4V(ision) System Card.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; Schuh, P.; Shi, K.; Tsvyashchenko, S.; Maynez, J.; Rao, A.; Barnes, P.; Tay, Y.; Shazeer, N.; Prabhakaran, V.; Reif, E.; Du, N.; Hutchinson, B.; Pope, R.; Bradbury, J.; Austin, J.; Isard, M.; Gur-Ari, G.; Yin, P.; Duke, T.; Levskaya, A.; Ghemawat, S.; Dev, S.; Michalewski, H.; Garcia, X.; Misra, V.; Robinson, K.; Fedus, L.; Zhou, D.; Ippolito, D.; Luan, D.; Lim, H.; Zoph, B.; Spiridonov, A.; Sepassi, R.; Dohan, D.; Agrawal, S.; Omernick, M.; Dai, A. M.; Pillai, T. S.; Pellat, M.; Lewkowycz, A.; Moreira, E.; Child, R.; Polozov, O.; Lee, K.; Zhou, Z.; Wang, X.; Saeta, B.; Diaz, M.; Firat, O.; Catasta, M.; Wei, J.; Meier-Hellstern, K.; Eck, D.; Dean, J.; Petrov, S.; and Fiedel, N. 2022. PaLM: Scaling Language Modeling with Pathways. arXiv:2204.02311.

Fernando, C.; Banarse, D.; Michalewski, H.; Osindero, S.; and Rocktäschel, T. 2023. Promptbreeder: Self-Referential Self-Improvement Via Prompt Evolution. arXiv:2309.16797.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. arXiv:2208.01618.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2018. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. arXiv:1706.08500.

Huang, K.; Sun, K.; Xie, E.; Li, Z.; and Liu, X. 2023. T2I-CompBench: A Comprehensive Benchmark for Open-world Compositional Text-to-image Generation. arXiv:2307.06350.

KandooAI. 2024. Juggernaut XL.

Khachatryan, L.; Movsisyan, A.; Tadevosyan, V.; Henschel, R.; Wang, Z.; Navasardyan, S.; and Shi, H. 2023. Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. arXiv:2303.13439.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2023. Large Language Models are Zero-Shot Reasoners. arXiv:2205.11916.

Ku, M.; Li, T.; Zhang, K.; Lu, Y.; Fu, X.; Zhuang, W.; and Chen, W. 2024. ImagenHub: Standardizing the evaluation of conditional image generation models. arXiv:2310.01596.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. arXiv:2201.12086.

Liang, C.; Ma, F.; Zhu, L.; Deng, Y.; and Yang, Y. 2024. CapHuman: Capture Your Moments in Parallel Universes. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6400–6409.

Liu, Y.; Zhang, K.; Li, Y.; Yan, Z.; Gao, C.; Chen, R.; Yuan, Z.; Huang, Y.; Sun, H.; Gao, J.; He, L.; and Sun, L. 2024. Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models. arXiv:2402.17177.

Nyberg, E. P.; Nicholson, A. E.; Korb, K. B.; Wybrow, M.; Zukerman, I.; Mascaro, S.; Thakur, S.; Oshni Alvandi, A.; Riley, J.; Pearson, R.; Morris, S.; Herrmann, M.; Azad, A.; Bolger, F.; Hahn, U.; and Lagnado, D. 2021. BARD: A Structured Technique for Group Elicitation of Bayesian Networks to Support Analytic Reasoning. *Risk Analysis*, 42(6): 1155–1178.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022a. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022b. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023a. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. arXiv:2208.12242.

Ruiz, N.; Li, Y.; Jampani, V.; Wei, W.; Hou, T.; Pritch, Y.; Wadhwa, N.; Rubinstein, M.; and Aberman, K. 2023b. Hy-

perDreamBooth: HyperNetworks for Fast Personalization of Text-to-Image Models. arXiv:2307.06949.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv:2205.11487.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved Techniques for Training GANs. arXiv:1606.03498.

Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. arXiv:2302.04761.

Shen, Y.; Song, K.; Tan, X.; Li, D.; Lu, W.; and Zhuang, Y. 2023. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. arXiv:2303.17580.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Wallace, B.; Dang, M.; Rafailov, R.; Zhou, L.; Lou, A.; Purushwalkam, S.; Ermon, S.; Xiong, C.; Joty, S. R.; and Naik, N. 2023. Diffusion Model Alignment Using Direct Preference Optimization. *ArXiv*, abs/2311.12908.

Wang, Z.; Zhao, L.; and Xing, W. 2023. StyleDiffusion: Controllable Disentangled Style Transfer via Diffusion Models. arXiv:2308.07863.

Wang, Z. J.; Montoya, E.; Munechika, D.; Yang, H.; Hoover, B.; and Chau, D. H. 2023. DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models. arXiv:2210.14896.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

Wu, C.; Yin, S.; Qi, W.; Wang, X.; Tang, Z.; and Duan, N. 2023. Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models. arXiv:2303.04671.

Yang, L.; Yu, Z.; Meng, C.; Xu, M.; Ermon, S.; and Cui, B. 2024a. Mastering Text-to-Image Diffusion: Recaptioning, Planning, and Generating with Multimodal LLMs. arXiv:2401.11708.

Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; Cui, B.; and Yang, M.-H. 2024b. Diffusion Models: A Comprehensive Survey of Methods and Applications. arXiv:2209.00796.

Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.-C.; Liu, Z.; and Wang, L. 2023a. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). arXiv:2309.17421.

Yang, Z.; Wang, J.; Li, L.; Lin, K.; Lin, C.-C.; Liu, Z.; and Wang, L. 2023b. Idea2Img: Iterative Self-Refinement with GPT-4V(ision) for Automatic Image Design and Generation. arXiv:2310.08541.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; Hutchinson, B.; Han, W.; Parekh, Z.; Li, X.; Zhang, H.; Baldridge, J.; and Wu, Y. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. arXiv:2206.10789.

Zhang, Y.; Huang, N.; Tang, F.; Huang, H.; Ma, C.; Dong, W.; and Xu, C. 2023. Inversion-Based Style Transfer with Diffusion Models. arXiv:2211.13203.

Zhang, Z.; Zhang, A.; Li, M.; and Smola, A. 2022. Automatic Chain of Thought Prompting in Large Language Models. arXiv:2210.03493.

Zhou, D.; Li, Y.; Ma, F.; Yang, Z.; and Yang, Y. 2024a. MIGC++: Advanced Multi-Instance Generation Controller for Image Synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–15.

Zhou, D.; Li, Y.; Ma, F.; Zhang, X.; and Yang, Y. 2024b. MIGC: Multi-Instance Generation Controller for Text-to-Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6818–6828.