Evaluation of Text-to-Image Generation from a Creativity Perspective

Xinhao Wang[1], Xinyu Ma[1], Shengyong Ding[2], Derek F. Wong[1]✉

[1]NLP[2]CT Lab, Department of Computer and Information Science, University of Macau
[2]Guangzhou DataStory Information Technology Co., Ltd.
[1]derekfw@um.edu.mo

Abstract

In recent years, driven by advancements in the diffusion process, Text-to-Image (T2I) models have rapidly developed. However, evaluating T2I models remains a significant challenge. While previous research has thoroughly assessed the quality of generated images and image-text alignment, there has been little study on the creativity of these models. In this work, we defined the creativity of T2I models, inspired by previous definitions of machine creativity. We also proposed corresponding metrics and designed a method to test the reliability of the metric. Additionally, we developed a fully automated pipeline capable of transforming existing image-text datasets into benchmarks tailored for evaluating creativity, specifically through text vector retrieval and the text generation capabilities of large language models (LLMs). Finally, we conducted a series of tests and analyses on the evaluation methods for T2I model creativity and the factors influencing the creativity of the models, revealing that current T2I models demonstrate a lack of creativity. The code and benchmark are available at https://github.com/pianopiece/T2ICreativity.

1 Introduction

Inspired by the diffusion process, researchers have designed a series of Text-to-Image (T2I) models, which exhibit outstanding performance and have significantly contributed to the development of image generation, such as Stable Diffusion (Rombach et al., 2022; Podell et al., 2023; Esser et al., 2024), FLUX (Labs, 2024) and DALL-E3 (Betker et al., 2023), demonstrating powerful capabilities in generating relevant visual images from textual input. Despite the rapid advancement of image generation, a significant challenge remains: automated image evaluation (Lin et al., 2025; Tu et al., 2024), where the primary focus is typically on image quality and text-image consistency. In contrast, the automated assessment of creative aspects in generated images has received relatively little attention.

In image quality evaluation, Inception Score (Salimans et al., 2016) has measured diversity with a pre-trained Inception network, while FID (Heusel et al., 2017) has compared the distribution of generated and real images. For text-image consistency, approaches typically have involved comparing generated captions with human-annotated ones (Hong et al., 2018), or utilizing the CLIP Score (Brooks et al., 2023; Li et al., 2024; Wu et al., 2023; Esser et al., 2024) with CLIP model (Radford et al., 2021), which quantifies the cosine similarity between image and text embeddings. T2I models have been capable of generating high-quality, stylistically distinct images, achieving high scores on existing evaluation metrics; however, the evaluation perspectives discussed above give limited attention to the creativity of the models. Evaluating creativity is crucial for measuring a model's ability to generate interesting content. This is especially important in assisting professionals in fields such as art, design, and innovation. At the same time, model creativity extends the practical value of the models, enabling it to contribute to the development of industries such as advertising, fashion, and entertainment. Karampiperis et al. (2014) has demonstrated that the creativity exhibited in text artifacts can be predicted using appropriate formulations of computational creativity metrics. Aghazadeh and Kovashka (2024)have defined the creativity of images as their uniqueness in advertisement image generation and have exhibited that current T2I models faced challenges when it comes to generating creative outputs and there was previously a lack of relevant evaluation metrics. However, current work on evaluating creativity has not defined the creativity of T2I models or designed corresponding metrics based on psychological or philosophical definitions of creativity. Building

✉Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 481–493
November 4-9, 2025

upon the broader definitions of machine creativity (Franceschelli and Musolesi, 2024) in previous works, we extended this concept to T2I models, providing a specific definition that comprises three components: **Value**, **Novelty**, and **Surprise**. **Value** refers to whether the images align to human's instruction. **Novelty** refers to the uniqueness of the image in relation to other images generated by the same model. The uniqueness of image refers to the content that is exclusively present in one image within a set generated by the same T2I model using the same prompt. This unique content may include aspects such as color, perspective, object composition and so on. **Surprise** refers to whether the images contain unexpected or surprising content.

Based on the definitions we proposed, we established corresponding metrics, benchmarks, and a pipeline capable of automatically generating benchmarks based on existing image-text datasets. The pipeline has the capability to create a benchmark where one prompt corresponds to multiple images by clustering and merging similar texts from text-image pairs. Through multiple experiments, we tested the proposed metrics and demonstrated their feasibility, which can accurately differentiate the creativity of various models based on the three dimensions mentioned above. Additionally, we explored various factors that influence the evaluation of model creativity. On the generated benchmark, we tested the creativity of different versions of Stable Diffusion and observed that while Value consistently increased with each version, surprisingly, both Novelty and Surprise did not follow the same upward trend and, in fact, showed a decline (up to -0.081, -0.019, respectively). This means that as the model upgrades, although it can produce high-quality images, the likelihood of generating imaginative, novel, and inspiring content has decreased, which is something we previously overlooked. This finding underscores the importance of evaluating model creativity.

In summary, the key contributions of our study are threefold:

1. Based on the general concept of machine creativity, we define the creativity of T2I models as consisting of Value, Novelty, and Surprise, and have designed evaluation methods along with relevant metrics.

2. We have designed a fully automated pipeline that can convert existing image-text datasets into the benchmark required for evaluating creativity, without the need for manual intervention.

3. We tested our proposed metrics and demonstrated their feasibility. Furthermore, we evaluated different T2I models on the generated benchmark and found that Novelty and Surprise did not increase with version updates; instead, they decreased. This highlights the importance of assessing creativity.

2 Related Works

2.1 T2I Models

T2I models based on the diffusion process soon gained widespread attention, leading to the emergence of numerous impressive models. Rombach et al. (2022) has presented a latent diffusion model, which significantly improved training efficiency and has the capability to generate high-quality, high-resolution images. Compared to previous versions of Stable Diffusion, Stable Diffusion XL (Podell et al., 2023) has designed a model with more parameters and introduced a refinement model to improve details. The model has achieved significant performance improvements over previous models. Stable Diffusion 3 (Esser et al., 2024) has improved existing noise sampling techniques and introduces a new transformer-based (Vaswani, 2017) model architecture, resulting in further performance enhancements.

2.2 T2I Metrics & Benchmarks

In recent years, designing automatic evaluation metrics to assess the quality of machine-generated images has always been a topic of great interest among researchers in the field of computer vision. Inception Score (Salimans et al., 2016) and Fréchet Inception Distance (Heusel et al., 2017) are the most widely adopted image quality metrics. The former has extracted visual features from generated images using a pre-trained Inception-V3 model (Szegedy et al., 2016) to evaluate image diversity. The latter has compared these extracted features with those of "gold" images to assess image fidelity. CLIPScore (Hessel et al., 2021) is based on computing the cosine similarity between image and text embeddings, as a metric for image-text alignment. VQAScore (Lin et al., 2025) has evaluated the alignment between an image and a text prompt by leveraging the latent knowledge of large models. It calculates the probability that the model

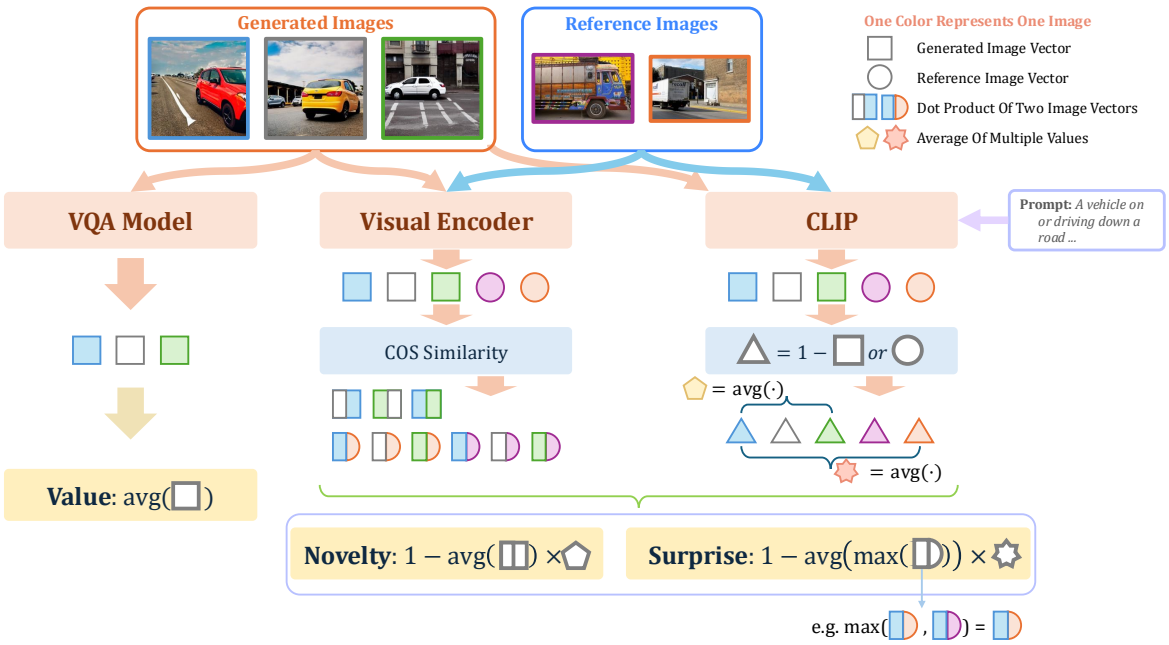

Figure 1: An illustration of the metric calculation process, including **Value**, **Novelty** and **Surprise**. Firstly, we encode the images with Visual encoder and compute the cosine similarity between the vectors of the images. Simultaneously, we calculate the text-image similarity by CLIP, which allows us to estimate the proportion of the visual semantics that lies outside the scope of the prompt. By using a weighted approach, we compute a more reasonable distance between the images to measure Novelty and Surprise. Additionally, we calculate the mean of the VQAScore as Value.

answers "Yes" to the question "*Does this figure show 'text'?*".

Additionally, a high-quality benchmark is urgently needed for evaluating T2I (Text-to-Image) models. Visual Genome (Krishna et al., 2017) and MSCOCO (Lin et al., 2014) are widely used datasets for computer vision research, consisting of large-scale real-world scenes annotated for tasks such as object detection, captioning and evaluating image quality and image-text consistency. TIFA v1.0 (Hu et al., 2023) is a benchmark that includes 4k diverse text inputs and 25k questions across 12 categories for T2I faithfulness evaluation. DSG-1k (Cho et al., 2023) has encompassed a broad spectrum of fine-grained semantic categories, ensuring a balanced distribution throughout.

3 Creativity Evaluation

3.1 Creativity Definition of T2I Model

Franceschelli and Musolesi (2024) considered Boden's criteria for studying machine creativity, which was first defined as "the ability to come up with ideas or artifacts that are new, surprising and valuable" (Boden, 2004). **Value** encompasses utility, performance, and attractiveness, and is connected to both the quality of production and its societal acceptance (Maher, 2010). **Novelty** refers to the degree of difference between the created artifact and others within its class (Ritchie, 2007). **Surprise** refers to how much a stimulus deviates from expectations (Berlyne, 1973).

In this work, we follow the multi-dimensional conception of machine creativity as outlined in the literature. Specifically, we have specified the three dimensions of value, novelty, and surprise as criteria for evaluating the creativity of T2I models. **Value** refers to whether the model can generate content that includes what is mentioned in the prompt. **Novelty** is used to evaluate whether the model can generate diverse images across multiple attempts using the same prompt. **Surprise** refers to whether the model has the ability to generate content in the image that is beyond expectations.

It is important to note that our work focuses on evaluating **the model's creativity**, rather than the creativity of the images generated by the model. **Value** is related to image-text consistency, but image-text consistency seeks a one-to-one correspondence between the image and the prompt. However, since the prompt is text, its information

is limited. We hope that the model, in addition to meeting the prompt's requirements, can demonstrate its own creativity, providing more inspiration for users and even artists. In such cases, a one-to-one correspondence is inappropriate from the perspective of creativity. Instead, it is more appropriate to judge whether the model has fulfilled the requirements of the prompt. As two metrics for evaluating creativity, **Novelty** and **Surprise** align more closely with people's intuitive understanding of creativity, being associated with originality and diversity. However, randomly generated images may also score high in novelty or surprise, for example, the images produced with guidance scale of 1 as shown in Fig 2. This is where **Value** plays a constraining role in our assessment of creativity: when the generation is arbitrary, value tends to be low, indicating that the creativity demonstrated by the model lacks meaningful contribution.

3.2 Creativity Metric

3.2.1 Value

To evaluate whether the images generated by a model effectively capture the content described in the prompts, we chose to use VQAScore as the evaluation metric. CLIP is trained via contrastive learning to establish a one-to-one correspondence between images and text. In contrast, VQAScore evaluates the likelihood of a "Yes" response from a Large Vision Language Model (LVLM) when queried with relevant questions. LVLMs are typically trained on large-scale datasets and support more flexible question forms, whereas CLIP is limited to calculating relatively rigid image-text similarity. We took the average VQAScore of a set of generated images as the score of Value for the model when generating this set of images, based on the following formula.

$$Value = \frac{1}{N}\sum_{n=1}^{N} \text{VQAScore}(i_n^g, t) \tag{1}$$

where i_n^g represents the n^{th} generated image, while t denotes the prompt for image generation, and N is the number of generated images.

3.2.2 Novelty

According to the definition, we aimed to evaluate whether there were significant differences between images generated multiple times by the same model under the same prompt. We measured the visual semantic distance between generated images with visual encoder, which serves as the basis for calculating Novelty. As shown in Fig. 1, we also calculated the average of the image-text similarity between the generated images and the prompt, approximating this as the proportion of the prompt's semantics represented within the visual semantics. This allowed us to derive the proportion of other semantics beyond those included in the prompt in the visual content. Since all the generated images include the content of the prompts, our evaluation focuses on assessing the content beyond the prompts, which is our primary focus of interest. Specifically, we aimed to evaluate the semantic distance of non-prompt content generated across a T2I model's multiple attempts for generation. By leveraging the semantic proportion, we approximated the similarity of the content outside the prompts. Finally, the average semantic distance of the content out of prompt is calculated as Novelty score by averaging the similarity scores.

$$d_n^g = Encoder(i_n^g) \tag{2}$$

$$Prop_{nov} = 1 - \frac{1}{N}\sum_{n=1}^{N} CLIP(i_n^g, t) \tag{3}$$

$$Novelty = 1 - \frac{2}{N^2 - N} * \sum_{n=1}^{N}\sum_{j=n+1}^{N} cos\_sim(d_n^g, d_j^g) * Prop_{nov} \tag{4}$$

(5)

where d_n^g represents the visual embedding of the n^{th} generated image, and $Prop_{nov}$ in Novelty denotes the estimated proportion of similarity for content outside the prompt.

3.2.3 Surprise

Similar to how we evaluated Novelty, we aimed to evaluate whether the images generated multiple times by the model under the same prompt could contain content that exceeds human cognitive inertia. The Surprise evaluation process is similar to Novelty, with two main differences. One difference is that we introduce a reference image set. As mentioned in section 3.1, the Surprise metric is designed to evaluate whether the imaginative content of an image generated by a T2I model is beyond common knowledge. The reference image set consists of real images that not only contain the prompt's content but also include common content associated with the prompt. The Surprise

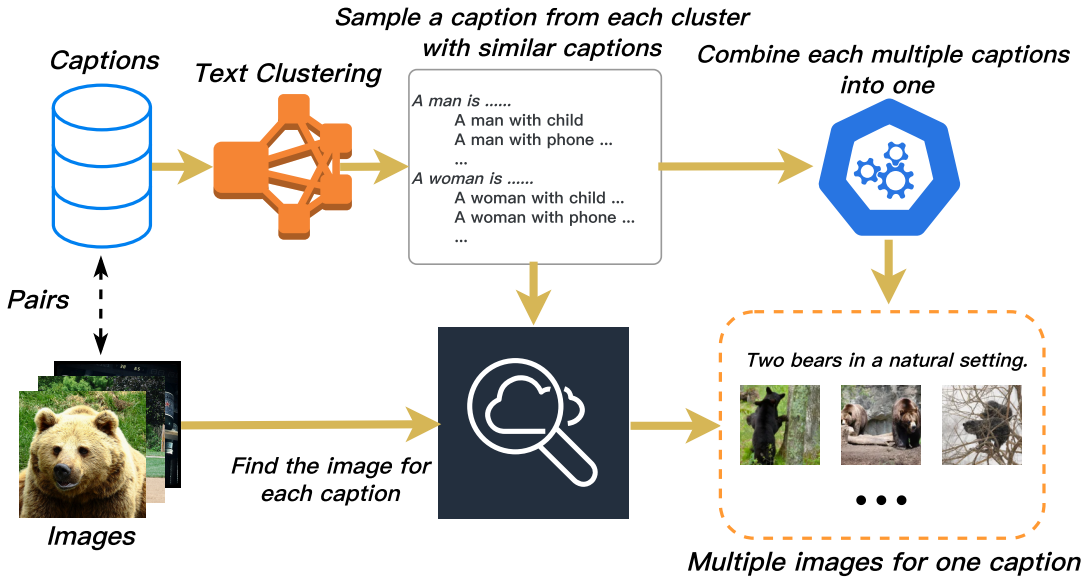

Figure 2: An illustration of fully automated benchmark generation pipeline. We constructed the benchmark for evaluating creativity through text clustering, the text summarization of LLMs, and existing image-text datasets. In this benchmark, each prompt is associated with multiple images.

is calculated by measuring the distance between the generated images and these reference images. Similarly to Novelty, we aimed to evaluate the distance among the contents beyond the prompt, as the prompt content is mandatory for all images. Therefore, we also introduced CLIP. The other difference is that, unlike Novelty, we selected the maximum similarity between a generated image and multiple reference images as the Surprise value for the T2I model under this prompt. The max similarity means the min Surprise. This is because our expectation for Surprise is more stringent; once the content is predictable, it is no longer a Surprise.

$$Prop_{surp} = 1 - \frac{1}{N+S}[\sum_{n=1}^{N} CLIP(i_n^g, t) + \sum_{s=1}^{S} CLIP(i_s^r, t)] \quad (6)$$

$$Surprise = 1 - \frac{1}{N} * \sum_{n=1}^{N} \max_{s \in S} cos\_sim(d_n^g, d_s^r) * Prop_{surp} \quad (7)$$

where i_s^r and d_s^r represent the s^{th} reference image and its visual embedding respectively, and S is the number of reference images.

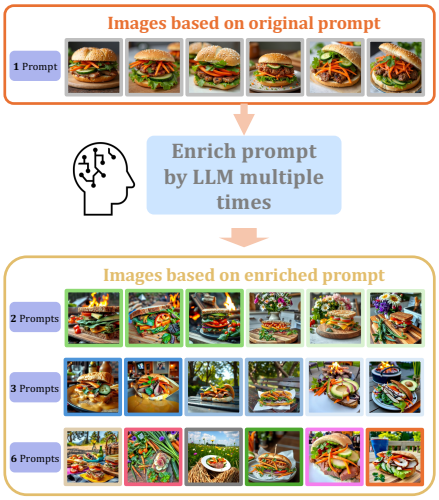

Figure 3: An illustration of the method for testing metric. Enhancing the content of original prompts through LLM while preserving their original semantics, aiming to enable the model to generate content that is richer compared to the original prompts.

3.3 Benchmark & Generation Pipeline

To evaluate the model's creativity, we constructed a fully automated process that can transform existing image-text datasets into benchmarks required for assessing creativity, as depicted in the Fig. 2.

By encoding the text in the image-text pairs of the dataset and then clustering them, all text vectors are divided into n categories, where n depends

on the desired size of the benchmark. Next, the pipeline randomly selects one text from each cluster and, based on similarity calculations, finds the $k-1$ most similar prompts within the same cluster. The value of k depends on the number of reference images needed for evaluating Surprise. Then retrieve the images corresponding to these k prompts to serve as reference images. Finally, the pipeline merges the k prompts into a single prompt with a LLM, ensuring that the merged prompt corresponds to all the reference images, with the prompt, "Here are some captions. '*{captions}*' Please find what these captions have in common, don't have to describe the difference between them, DO NOT use generalisations such as various, different and so on and write it in one caption. Please only answer the caption without anything else.". In this paper, the value of k is 6, resulting in a benchmark consisting of 384 prompts and their reference images based on MSCOCO (Lin et al., 2014).

4 Experiments

4.1 Test for Metric

Through extensive experiments and consistency tests with human judgments, Fu et al. (2024) found that the DINO model is capable of capturing subtle differences in visual semantics. Therefore, we chose the DINOv2 large model (Oquab et al., 2023) as the visual encoder. For the Value metric, we directly used the VQAScore, so no additional testing is required. In our subsequent test experiments, we used the FLUX API provided by Alibaba to generate high-quality images for testing.

We designed a method, illustrated in Fig. 3, to test whether the Novelty metric can distinguish between image sets with different levels of Novelty. For evaluation, we set the T2I model to run six times to generate six different images. We predefined four levels of Novelty image sets, ranging from low to high, using an original prompt, two enriched prompts, three enriched prompts, and six enriched prompts, respectively, and two levels of Surprise image sets, ranging from low to high, using an original prompt, and all the other enriched prompts, respectively. For a detailed explanation of the method in the figure, please refer to the appendix A.

We selected 100 prompts from TIFA benchmark (Hu et al., 2023). As shown in Fig. 4, the ranking of the results evaluated by the Novelty metric aligns with our predefined ranking, from low to high, one prompt, two prompts, three prompts, and six prompts, respectively. This demonstrates that our metric can distinguish the rankings of image sets with different levels of Novelty which is predefined. As expected, the other image sets have significantly higher Surprise values compared to the 1 prompt image set, while the Surprise values among the other image sets are similar. In summary, our defined metrics can distinguish between the levels of Novelty and Surprise as defined in the previous section.

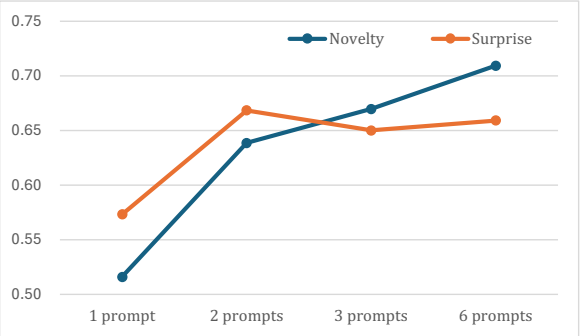

Figure 4: Test results for Novelty and Surprise. As the number of enriched prompts increases, Novelty also gradually rises. Additionally, Surprise is significantly enhanced when comparing image sets generated with enriched prompts to those generated with a single prompt. The aforementioned results align with our expectations.

Model	Value↑	Novelty↑	Surprise↑
SD-v1-4	0.7858	0.5792	0.6232
SD-XL	0.8080	0.5511	0.6212
SD-v3med	0.8283	0.4981	0.6040
Kolors	0.7982	0.4639	0.6284
CogView4	0.8035	0.4723	0.6440
Janus-Pro	0.8158	0.4391	0.6137
BLIP3o 4B	0.8323	0.4003	0.6015

Table 1: Experimental results on benchmark. Value, has gradually increased with model update. However, in the context of creativity, the newly introduced metrics of Novelty and Surprise show the opposite trend.

4.2 Results on Benchmark

As shown in Table 1, the value increases with the update of stable diffusion versions. This indicates that the model is increasingly able to accurately generate content that includes the prompt, aligning with the expected model improvements. However, under the Novelty and Surprise metrics, the situation is the opposite, especially for Novelty. The

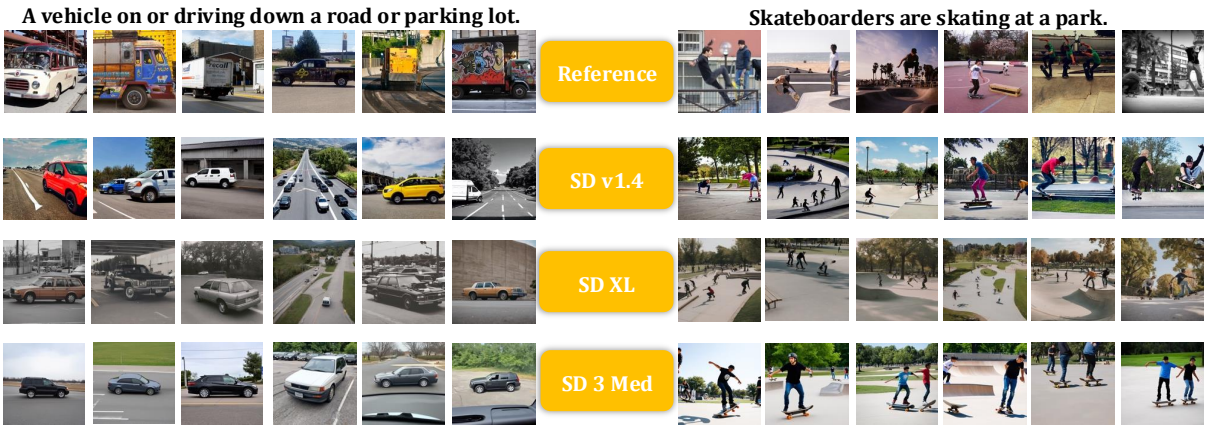

Figure 5: A generation case of benchmark. Stable Diffusion v1.4 demonstrates considerable variation in the generated images. Conversely, Stable Diffusion v3 medium exhibits minimal variation, maintaining a consistent visual angle and color palette for the car, as well as uniformity in the depiction of skateboarders. This suggests that when evaluating model performance, creativity was rarely considered before.

decrease in Novelty (up to -0.081) means that the content generated by the model tends to become more homogeneous over multiple generations and in Surprise (up to -0.019) indicates that the content generated by the model becomes less likely to exceed people's expectations. It is clear, based on Fig. 5, that in the generation tasks of these two prompts, Stable Diffusion v1.4 exhibits significant variation in color schemes, visual angles, and compositional elements across multiple generations. In stark contrast, Stable Diffusion v3 medium shows little variation, with the visual angle and color of the car remaining largely consistent, and the content related to skateboarders following the same pattern. This means that in the past, we did not focus much on improving model performance from the perspective of creativity. The specific experimental parameters are provided in the appendix B. Meanwhile, we demonstrated through significance testing that the performance decline on novelty and surprise metrics is statistically significant. The significance testing results are provided in the appendix C. Except for SD v1.4 and SD XL, although there is no significant difference in surprise, this aligns with our observations from the metric results, as the two models exhibit similar performance on surprise. At the same time, we also tested other series of text-to-image models, such as Kolors (Team, 2024), CogView4 (Zheng et al., 2024), Janus-Pro (Chen et al., 2025b), and BLIP3o 4B (Chen et al., 2025a). From the results, we observed that although the generated images aligned well with the prompts and did not depict the common scenes seen in the reference images, they tended to be consistent across multiple generations, indicating low novelty. This suggests that multiple attempts to generate images with these models may not provide users with more inspiration or reference.

	Value↑	Novelty↑	Surprise↑
Baseline	0.7858	0.5792	0.6232
w/ diff seeds	0.7854	0.5849	0.6271
w/ 20 images	0.7863	0.5749	0.6249
w/ gs 12.5	0.7872	0.5645	0.6240
w/ gs 5	0.7782	0.6025	0.6290
w/ gs 1	0.5707	0.7801	0.7749

Table 2: Experimental results on the impact of the number of images and random seeds on the evaluation, and the effect of guidance scale on the model's creativity. gs refers to guidance scale.

4.3 Influential Factors Analysis

In this section, we analyze the impact of the number of generated images and different text to express the same prompt semantic on the evaluation of creativity, the effect of the guidance scale on the model's creativity. The guidance scale in T2I models controls how closely the generated image follows the text prompt. Higher values make the image adhere more strictly to the prompt, while lower values allow for more creative freedom. We choose the Stable Diffusion v1.4 that showed the best Novelty and Surprise performances in the benchmark experiment as the base model.

From the experimental results in Table 2, we can see that changing the random seed to generate

images six times again has a negligible impact, and generating more images to evaluate the model's creativity has little effect as well. This indicates that generating six images is sufficient to evaluate model creativity, and the performance is minimally affected by the random seed.

The default guidance scale of Stable Diffusion v1.4 (Rombach et al., 2022), is 7.5. Increasing guidance scale prompts the model to produce images that are more closely aligned with the text prompt. In our experiments, we tested the results with scales of 12.5, 5, and 1, keeping other parameters constant. We observed that appropriately lowering the guidance scale can increase Novelty while keeping the value relatively unchanged, with a slight fluctuation in Surprise (up to -0.007 Value, +0.023 Novelty and +0.006 Surprise). However, if the guidance scale is reduced to 1, although both Novelty and Surprise increase significantly, the Value drops sharply. This indicates that the high Novelty and Surprise are due to the image content deviating from the prompt, as shown in Appendix D.

We sampled 50 prompts from the benchmark and used an LLM to rewrite the prompts in each group into different expressions without changing the semantics, with the prompt, "Here is a caption. '*{caption}*'. Please rewrite this caption without changing the meaning of the sentence and only answer the rewritten caption directly without anything else.". Each prompt was rewritten twice, resulting in a total of three versions including the original prompt. Each prompt generated two images, totally six images. From Table 3, we can find that the expression of the prompt has a minimal impact on evaluating the model's creativity under the same semantics. This result also indicates that simply altering the form of the prompt is not a feasible approach to enhancing model creativity.

	Value↑	Novelty↑	Surprise↑
Baseline	0.7665	0.5967	0.6503
w/ rewrite	0.7684	0.6023	0.6399

Table 3: Experimental results on the effect of prompt expression on evaluation. The prompts were rewrote by LLM without altering their semantics.

4.4 Metric Analysis

We randomly selected 50 samples from the generated images of three StableDiffusion series models and conducted a human preference alignment test for Novelty and Surprise. Human annotators ranked the three models based on the definitions of novelty and surprise. As shown in Table 4, the rankings based on our metrics, Novelty and surprise, exhibited an average Pairwise Accuracy of 0.71 and 0.61, respectively, and a Hit@1 of 0.54 and 0.44, respectively. Although the two metrics have not yet reached the level of consensus observed among humans, they demonstrate good consistency with human preferences, providing a reliable and robust indicator for evaluating creativity.

In Table 1, we observed a potential correlation between the values of Novelty and Surprise. To further investigate this relationship, we computed the correlation between the novelty and surprise rankings for each sample. As shown in Table 4, the rankings of the two metrics exhibit a certain degree of association. Based on the results of human preference alignment, theoretical definitions, and computational methods, we hypothesize that this correlation arises from the limited precision in the assessment of surprise. A more direct cause lies in the insufficient coverage of reference images within the benchmark. However, datasets containing a large number of images per prompt are currently scarce. From a definitional perspective, it is intuitive to distinguish between the two metrics: Novelty emphasizes that the model should generate images with unique features each time, whereas Surprise focuses on the model's ability to produce content that is new relative to what is already known by humans.

	Pairwise Accuracy↑	Hit@1↑
H, N	0.71	0.54
H, S	0.61	0.44
N, S	0.61	0.46

Table 4: Correlation among human annotators, Novelty and Surprise. H, N, S refers to human, Novelty and Surprise, respectively.

5 Conclusion

In this paper, we explore the definition of creativity and its application in T2I models. For evaluation, we propose creativity metrics, consisting of Value, Novelty and Surprise, and a fully automatic benchmark generation pipeline. Experimental results across the generated benchmark validate creativity is a new, valuable perspective for T2I model evalu-

ation. Furthermore, we conducted detailed analysis experiments on the influences of hyper-parameters on the evaluation of creativity.

Limitations

Despite the contributions of this work, there are several limitations that should be acknowledged. The limitations define the boundaries of our current work and suggest directions for future research.

1. When assessing the impact of the same set of images with identical semantics on the evaluation of Novelty and Surprise, we employed CLIP to approximate the semantic proportion and evaluate the distance between other semantics in different images in the set, excluding those with identical semantics. However, this method is not entirely appropriate, and a more precise approach is needed to measure the semantics we intend to compare.

2. This work focuses on evaluating the creativity of the model. For assessing the creativity of a single image, current methods may not be entirely suitable. A larger and more diverse image dataset might be necessary to support image creativity evaluation. Additionally, creative elements such as metaphors embedded within a single image may require deep exploration by large language models to be better evaluated.

Ethical Considerations

Our benchmark is derived from MSCOCO, which is licensed under the Creative Commons Attribution 4.0 License. Dinov2 large is distributed under the Apache License 2.0, while CLIP ViT-Large Patch 14 adheres to the MIT License. LLaVA 1.5 is governed by the LLAMA 2 Community License.

Our usage of these models and benchmarks in this study is strictly for academic purposes and follows license.

Acknowledgements

This work was supported in part by the Science and Technology Development Fund of Macau SAR (Grant No. FDCT/0007/2024/AKP), the Science and Technology Development Fund of Macau SAR (Grant No. FDCT/0070/2022/AMJ, China Strategic Scientific and Technological Innovation Cooperation Project Grant No. 2022YFE0204900), the Science and Technology Development Fund of Macau SAR (Grant No. FDCT/060/2022/AFJ, National Natural Science Foundation of China Grant No. 62261160648), the UM and UMDF (Grant Nos. MYRG-GRG2023-00006-FST-UMDF, MYRG-GRG2024-00165-FST-UMDF, EF2024-00185-FST), and the National Natural Science Foundation of China (Grant No. 62266013).

References

Aysan Aghazadeh and Adriana Kovashka. 2024. Cap: Evaluation of persuasive and creative image generation. *arXiv preprint arXiv:2412.10426*.

Daniel E Berlyne. 1973. Aesthetics and psychobiology. *Journal of Aesthetics and Art Criticism*, 31(4).

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

Margaret A Boden. 2004. *The creative mind: Myths and mechanisms*. Routledge.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. 2025a. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*.

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

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, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*.

Giorgio Franceschelli and Mirco Musolesi. 2024. Creativity and machine learning: A survey. *ACM Computing Surveys*, 56(11):1–41.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. 2024. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *Advances in Neural Information Processing Systems*, 36.

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

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

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. 2018. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7986–7994.

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*, pages 20406–20417.

Pythagoras Karampiperis, Antonis Koukourikos, and Evangelia Koliopoulou. 2014. Towards machines for measuring creativity: The use of computational tools in storytelling activities. In *2014 IEEE 14th International Conference on Advanced Learning Technologies*, pages 508–512. IEEE.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Black Forest Labs. 2024. Flux. https://github.com/black-forest-labs/flux.

Dongxu Li, Junnan Li, and Steven Hoi. 2024. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36.

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. 2025. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Mary Lou Maher. 2010. Evaluating creativity in humans, computers, and collectively intelligent systems. In *Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design*, pages 22–28.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

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, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Graeme Ritchie. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines*, 17:67–99.

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.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

K Team. 2024. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*.

Rong-Cheng Tu, Zi-Ao Ma, Tian Lan, Yuehao Zhao, Heyan Huang, and Xian-Ling Mao. 2024. Automatic evaluation for text-to-image generation: Task-decomposed framework, distilled training,

and meta-evaluation benchmark. *arXiv preprint arXiv:2411.15488*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633.

Wendi Zheng, Jiayan Teng, Zhuoyi Yang, Weihan Wang, Jidong Chen, Xiaotao Gu, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. Cogview3: Finer and faster text-to-image generation via relay diffusion. In *European Conference on Computer Vision*, pages 1–22. Springer.

A Explanation of the method for testing metric

We enriched a prompt through LLM while retaining its original semantics, with the LLM prompt, *"image caption: {caption}*
nPlease expand the image caption to provide more elements that are not present in the caption, even change a different environment. Please note, however, that the rewrited caption must include the original content." By altering the prompt, we forced the T2I model to generate images containing the original prompt content in different scenarios. If we generate six images with an original prompt, these images will be quite similar. However, if the model generates with six enriched prompts, each generating one image, these six images will be significantly different. With two enriched prompts, each generating three images, results in six images with two groups, where the images between the groups are more different, and those in the same group more similar. Similarly, using three enriched prompts follows the same logic.

The essence of the evaluation process for Novelty and Surprise is fundamentally consistent, with the main difference being that Novelty involves comparing generated images with each other, while Surprise involves comparing them with references. To test the Surprise metric, the model generated two images with the original prompt serving as reference images. But it is not possible to pre-set levels for Surprise. It is hard to control Surprise by adjusting the number of enriched prompts as we did with Novelty. Novelty involves comparing generated images with each other, where controlling the enriched prompts ensures that images generated under the same enriched prompt are similar, while images generated under different enriched prompts are significantly different. However, Surprise involves comparing the generated images with the reference images which are fixed. As long as the images generated from the enriched prompts are significantly different from the reference images provided by the original prompt, we could only preset this one ranking, i.e., the 2, 3, and 6 prompts image sets would rank higher than the 1 prompt image set. However, we could not preset the rankings among the 2, 3, and 6 prompts image sets.

B Implementation Details for Benchmark

We conducted the experiments on three typical T2I models: Stable Diffusion v1.4 (Rombach et al., 2022), Stable Diffusion XL base 1.0 (Podell et al., 2023), Stable Diffusion 3 medium (Esser et al., 2024), Kolors (Team, 2024), CogView4 (Zheng et al., 2024), Janus-Pro (Chen et al., 2025b), and BLIP3o 4B (Chen et al., 2025a). For the visual encoder, as stated in the previous section, we selected the DINOv2 large model (Oquab et al., 2023). For the CLIP model, we chose to use CLIP ViT-Large Patch 14 created by OpenAI. We ran the experiments on a single RTX 4090D. All models output at default resolutions. Specifically, the output resolutions for Stable Diffusion v1.4, XL, and 3, Kolors, and CogView4 are 512x512, 1024x1024, 1024x1024, 1024x1024 and 512x512, respectively. For calculating VQAScore, we chose LLaVA v1.5 7B (Liu et al., 2024) as the base model, following Lin et al. (2025). The number of inference steps and guidance scale (Ho and Salimans, 2022) are default, which are guided by the official repository documents on Huggingface.

C Statistical Significance of Evaluation Metrics among the StableDiffusions

Table 5 provides statistical significance of evaluation metrics among the stableDiffusion model series.

D Different Guidance Scale Results

Image Table 6 provides detailed generated images on benchmark with different guidance scale setting.

E Annotator Information and Ethical Considerations

Four human annotators participated in the data labeling process. All annotators held at least a bachelor's degree and came from diverse academic backgrounds, ensuring a breadth of perspectives in the annotation. They volunteered their time, as part of an academic collaboration. Prior to participation, all annotators provided informed consent for the use of the data in this study. The instructions provided to the annotators included the definitions of "Novelty" and "Surprise" as described in the context of this work, which guided their labeling tasks.

	Novelty		Surprise	
	t-statistic	p-value	t-statistic	p-value
SD-v1-4, SD-XL	6.33	6.95E-10	0.57	0.57202
SD-v1-4, SD-v3med	15.52	1.80E-42	5.01	8.31E-07
SD-XL, SD-v3med	10.30	4.15E-22	4.57	6.50E-06

Table 5: Statistical significance test results between different models. Model scores are computed based on Novelty and Surprise metrics over 384 samples. The table reports t-statistics and p-values from paired t-tests between model pairs. A p-value below 0.05 indicates statistically significant performance differences.

Table 6: Cases of images generated under different guidance scale. It is evident that appropriately reducing guidance scale can enrich image content, thereby enhancing Novelty and Surprise. However, excessively lowering guidance scale, while significantly boosting Novelty and Surprise, results in images that are irrelevant to the prompt.